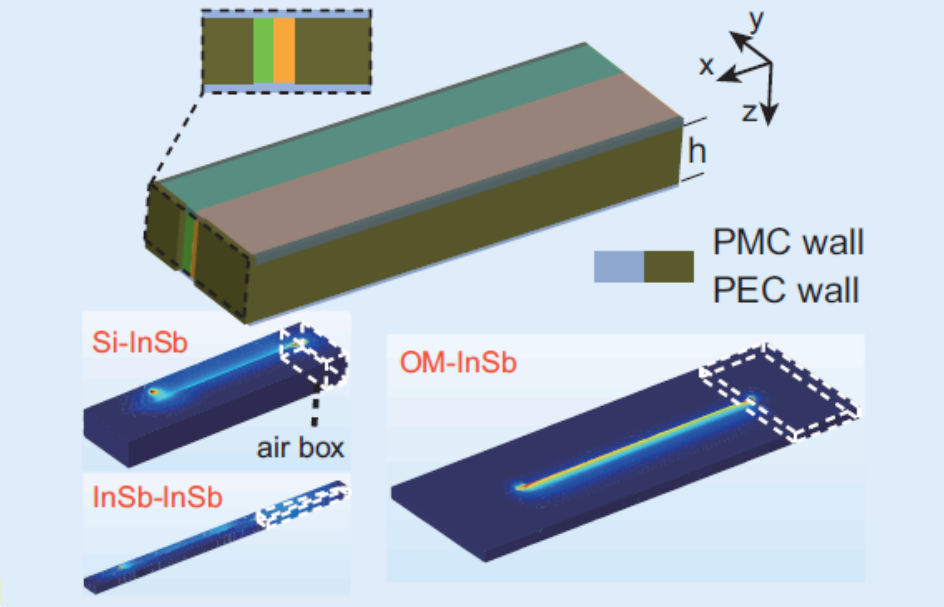
NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
DEPARTMENT OF PHYSICS
SECTION OF CONDENSED MATTER PHYSICS

PhD Thesis

Light transmission through holes in the deep subdiffractional regime using topological structures

Konstantinos G. Baskourelos

supervisor : Kosmas L. Tsakmakidis, Assist. Prof.



Athens, November 2023

Contents

Περίληψη στην ελληνική

Το θέμα της παρούσας εργασίας είναι μια καινοτόμα τεχνική για τη μετάδοση και εστίαση ενέργειας στη μικρο- και νανοκλίμακα, ειδικότερα για τη μετάδοση φωτός διαμέσου οπών ή σχισμών στο βαθύ υποπεριθλαστικό καθεστώς (δηλ. σημαντικά μικρότερων του μήκους κύματος λειτουργίας). Αυτή η τεχική έχει πολλά σημαντικά πλεονεκτήματα έναντι των συμβατικών τεχνικών για τον ίδιο σκοπό, και θα μπορούσε να επηρεάσει θετικά την ανάπτυξη του ευρύτερου κλάδου των νανοφωτονικών εφαρμογών.

Η μετάδοση και εστίαση φωτός στη μικρο- και νανοκλίμακα είναι στον κορμό πολλών σύγχρονων εφαρμογών : οπτική εγγραφή/αποθήκευση δεδομένων, θερμικά υποβοηθούμενη μαγνητική εγγραφή (HAMR[1]), νανοεικονοσκόπηση, φασματοσκοπία, αισθητήρες, οπτική ή θερμική νανοσκοπία κοντινού πεδίου, λιθογραφία με πρόμπες θερμικής ανίχνευσης, θερμομετρία νανοκλίμακας, και άλλες όπως αυτές. Σε όλες αυτές τις εφαρμογές, υπάρχει η απαίτηση να εστιαστεί με υψηλή απόδοση ισχύς ~ 100 μW σε μια περιοχή ~ 10 nm (ή μικρότερη) μιας επίπεδης επιφάνειας. Αυτή είναι μια ένταση φωτός εξαιρετικά υψηλή, πολλές τάξεις μεγέθους μεγαλύτερη από τη φωτεινή ένταση που συναντάται στα καθημερινά φυσικά φαινόμενα (π.χ., την ένταση του ηλιακού φωτός στην επιφάνεια της γης, ή τη φωτεινή ένταση που επιτυγχάνεται με έναν οπτικό φακό). Η Οπτική θέτει ένα κατώφλι στο φως που μπορεί να μεταδοθεί και να εστιαστεί από μια οπή δοθείσης διαμέτρου (σχέση του Bethe [22]) ˙ αυτό το κατώφλι καθιστά τις προαναφερθείσες εφαρμογές πολύ δύσκολο να πραγματοποιηθούν. Η καθιερωμένη τεχνολογία για την εστίαση του φωτός στη νανονκλίμακα είναι οι κωνικές οπτικές ίνες με επικάλυψη χρυσού, ευρέως χρησιμοποιούμενες στα Οπτικά Μικροσκόπια Ανίχνευσης Κοντινού Πεδίου (NSOMs[2]). Η απόδοση της οπτικής μετάδοσης του άκρου μιας NSOM πρόμπας είναι τυπικά μεταξύ 10^{-5} - 10^{-4} (ή μικρότερη). Ακόμη και η επίδοση των lasers είναι μακριά από την απόδοση της μετάδοσης που απαιτείται εδώ[3].

Η προκείμενη προταθείσα τεχνική, καλούμενη APOTUS-HM[4], είναι ένας τρόπος να ξεπεραστεί το κατώφλι που τίθεται από τη σχέση του Bethe. Η APOTUS-HM παρέχει έναν συντελεστή μετάδοσης ασύγκριτα μεγαλύτερο από άλλες καθιερωμένες τεχνικές, ο οποίος ιδεατά προσεγγίζει τη μονάδα ˙ ταυτόχρονα η βασική της ιδέα είναι απλή και αρκετά εύκολο να υλοποιηθεί στην πράξη. Η APOTUS-HM θεμελιώνεται σε τρεις βασικούς πυλώνες : (i) μονοκατευθυντικότητα στη διάδοση, (ii) προστασία από τη δια-

[1] HAMR : Heat-Assisted Magnetic Recording

[2] NSOM : Near field Scanning Optical microscope

[3] Π.χ., ένα φτηνό laser διόδου 10 mW έχει αποδοτικότητα μετάδοσης $\sim 1\%$ για μια περιοχή $d \geq 200$ nm ˙ ωστόσο, αυτή εξακολουθεί είναι μια πολύ μεγάλη διάμετρος για την απαιτούμενη στις περιπτώσεις που αναφέρθηκαν εδώ.

[4] **APOTUS-HM :** Almost Perfect Optical Transmission through Unstructured Single Hole Method

σπορά, και (iii) εξαιρετική οπτική μετάδοση (EOT[5]). Η βασική αρχή της APOTUS-HM είναι ως ακολούθως. Χρησιμοποιώντας ειδικά υλικά επιβάλλεται σε ένα κύμα να διαδοθεί σε έναν κυματοδηγό μόνον έμπροσθεν, μονοκατευθυντικά. Στο τέλος του κυματοδηγού υπάρχει μια οπή με την κατάλληλη διάμετρο για την εστίαση. Όταν το κύμα φτάσει στο πέρας του κυματοδηγού, καθώς δεν μπορεί να κινηθεί προς τα πίσω, εξαναγκάζεται να διέλθει από την οπή, ανεξάρτητα από το πόσο μικρή είναι, και άρα να εστιαστεί μπροστά της.

Παρά την απλότητα της παραπάνω ιδέας, για να υλοποιηθεί επιτυχώς μια συσκευή APOTUS-HM υπάρχουν πολλά λεπτά θεωρητικά θέματα που πρέπει να μελετηθούν και να κατανοηθούν. Στην προκείμενη εργασία γίνεται μια προσπάθεια να παρουσιαστεί πώς να εφαρμοστούν αποτελεσματικά τα τρία θεμέλια που αναφέρθηκαν παραπάνω και η βασική θεωρία πίσω από αυτά, δίνοντας με αυτόν τον τρόπο ένα υπόβαθρο για την καλύτερη κατανόηση και περαιτέρω βελτίωση της τεχνικής˙ και βέβαια δίνονται λεπτομέρειες αριθμητικών προσομοιώσεων βασικών συσκευών (μοντέλων) για την τεχνική.

Η μονοκατευθυντικότητα και η προστασία της διάδοσης από διασπορά επιτυγχάνεται χρησιμοποιώντας τοπολογικά υλικά. Το πρώτο και δεύτερο κεφάλαιο είναι μια γενική εισαγωγή στα τοπολογικά υλικά και σε μερικές πολύ σημαντικές παραμέτρους που τα χαρακτηρίζουν. Ειδικότερα, εισάγονται και συζητώνται εκτενώς η φάση Berry και οι αριθμοί Chern. Παρουσιάζονται εν συντομία μέθοδοι για τον αριθμητικό τους υπολογισμό. Εισάγεται η ιδέα του τοπολογικού υλικού, και μεταξύ άλλων συζητείται η σχέση της με τη χρονική συμμετρία και την Αρχή Ανταπόκρισης Όγκου-Ακμής (Bulk-Edge Correspondence Principle).

Το τρίτο κεφάλαιο είναι μια σύνοψη διαφόρων εφαρμογών – κάποιες εκ των οποίων πολύ ενδιαφέρουσες – που έχουν τα τοπολογικά υλικά στη Φωτονική. Η APOTUS-HM θα μπορούσε να βελτιώσει την απόδοση σε πολλές εξ αυτών.

Το τέταρτο κεφάλαιο είναι αφιερωμένο στη εξαιρετική οπτική μετάδοση (EOT). Η EOT είναι ένα κρίσιμο φαινόμενο για την APOTUS-HM καθώς λαμβάνει χώρα εκτενώς σε αυτή και αυξάνει τον συντελεστή μετάδοσης. Παρόλο που η EOT έχει ερευνηθεί εκτενώς σε πολλές μελέτες, εν προκειμένω είναι χρήσιμη μια ανασκόπηση των βασικών της χαρακτηριστικών. Επίσης, αναπαρήχθησαν μερικά αποτελέσματα της EOT για να αποκτηθεί διαίσθηση και να εκτιμηθεί καλύτερα ο ρόλος της στην APOTUS-HM.

Στο πέμπτο κεφάλαιο τέλος, εισάγεται και μελετάται η APOTUS-HM σε όλες τις πτυχές της. Παρουσιάζονται οι βασικές της ιδέες, και συζητώνται μοντέλα δομών που μπορούν να υλοποιηθούν στην πράξη. Καταρχήν, παρουσιάζονται εν συντομία μερικές ιδιότητες των επιφανειακών μαγνητοπλασμονίων (SMPs[6]), καθώς τα SMPs είναι άλλο ένα βασικό συστατικό της προκείμενης τεχνικής και δεν είναι τόσο γνωστά όσο τα επιφανειακά πλασμονικά πολαριτόνια (SPPs[7]). Αναπτύσσεται μια υποστηρικτική θεωρία[8], μικρή και απλή, αλλά αρκετά "έξυπνη", η οποία αφορά τη χρονική σύζευξη των ρυθμών. Αυτή η θεωρία δείχνει ότι η μετάδοση με την APOTUS-HM είναι επί της αρχής ανεξάρτητη από το πόσο μικρή είναι η οπή (μόνον οι απώλειες και η θέση της οπής παίζουν ρόλο). Σε μερικές δομές της APOTUS-HM είναι απαραίτητη η χρήση τέλειου *μαγνητικού*

[5] EOT : Extraordinary Optical Transmission

[6] SMP : Surface MagnetoPlasmon

[7] SPP : Surface Plasmon Polariton

[8] Αυτή η μικρή θεωρία είναι ουσιαστικά η θεωρητική θεμελίωση πώς ξεπερνάται το κατώφλι του Bethe.

αγωγού (PMC[9]) ως επικάλυψη του κυματοδηγού, σε αντίθεση με τον τέλειο ηλεκτρικό αγωγό (PEC[10]) που χρησιμοποιείται συνήθως. Οι PMCs δεν υπάρχουν στη φύση, έχουν ειδικές ιδιότητες και η υλοποίησή τους δεν είναι τετριμμένη[11]. Για την πληρότητα του κειμένου, μια ενότητα αυτού του κεφαλαίου είναι αφιερωμένη στις βασικές ιδιότητες των PMCs. Παρουσιάζονται προσομοιώσεις δομών που μπορούν να χρησιμοποιηθούν για την υλοποίηση συσκευών APOTUS-HM και μελετώνται οι βασικές τους ιδιότητες που αφορούν τη μετάδοση, διάδοση και διασπορά, σε 2D και 3D περιπτώσεις.

Συμπερασματικά, τα αποτελέσματα είναι πολύ ενθαρρυντικά και δείχνουν ότι η υπέρβαση του κατωφλίου Bethe με την APOTUS-HM είναι εφικτή, και δίχως κανένα σοβαρό περιορισμό επί της αρχής. Σύμφωνα με την αναπτυχθείσα υποστηρικτική θεωρία, η μετάδοση μέσω μιας μικρής οπής με την APOTUS-HM δεν εξαρτάται από το πόσο υποπεριθλαστική είναι η οπή, και έχει έναν συντελεστή μετάδοσης που πλησιάζει ακόμη και τη μονάδα. Αυτό το αποτέλεσμα είναι αξιοσημείωτο και ως τώρα δεν έχει ποτέ παρατηρηθεί ή επιτευχθεί στον κλάδο της Φωτονικής, ακόμη και με τις καλύτερες διαθέσιμες τεχνικές (χρήση SPPs και EOT φαινόμενα, κωνικές οπτικές ίνες κλπ) όπου η μετάδοση στο βαθύ υποπεριθλαστικό καθεστώς είναι πρακτικά αμελητέα[12]. Συνεπώς, παρόλο που πολλά απομένει να γίνουν για την ωριμότητα της τεχνικής, όπως σημειώθηκε νωρίτερα η APOTUS-HM μπορεί να έχει μια σημαντική επίδραση στον ευρύτερο κλάδο της Φωτονικής (EOT και τοπολογικοί κυματοδηγοί) και των εφαρμογών της στη μίκρο- και νανοκλίμακα γενικότερα.

Νοέμβριος 2023, Αθήνα,

Κωνσταντίνος Μπασκουρέλος

---

[9] PMC : Perfect Magneti Conductor

[10] PEC : Perfect Electric Conductor

[11] Η χρήση των PMCs, οπουδήποτε χρειάζεται, είναι πιθανόν ένα από τα ελάχιστα μειονεκτήματα της τεχνικής. Ωστόσο, επισημαίνεται ότι σε πολλές από τις εξεταζόμενες δομές η χρήση των PMCs δεν υφίσταται καν.

[12] Για παράδειγμα, με συμβατικές τεχνικές, μετάδοση από οπές $\sim \lambda_{eff}/50$ έχει συντελεστή μετάδοσης της τάξης 10^{-4} ή μικρότερο.

Abstract

The subject of the herein thesis is a novel technique for the transmission and focusing energy in micro- and nanoscale, particularly for the transmission of light through holes or slits in the deep subdiffractional regime (i.e., significantly smaller than the operating wavelength). This technique has many important advantages versus the conventional techniques for the same task, and it could affect positive the evolving of the wider field of the nanophotonic applications.

The transmission and focusing of light in the micro- and nanoscale is in the core of many contemporary applications : optical data writing/storage, heat-assisted magnetic recording (HAMR), nanoimaging, spectroscopy, sensing, near-field scanning optical or thermal nanoscopy, thermal scanning probe lithography, nanoscale thermometry, and others such these. In all these applications, there is the requirement to focus with high efficiency ~ 100 μW of power to a ~ 10 nm (or less) spot on a planar surface. This is a light intensity extremely high, many orders of magnitude larger than the light intensity encountered in everyday physical phenomena (e.g., the intensity of sunlight on the surface of the earth, or the light intensity attained with an optical lens). Optics poses a threshold to the light that can be transmitted and focused from a hole of given diameter (Bethe's relation [22]) ˙ this threshold makes the aforementioned applications very difficult to realize. The standard technology attaining the focusing of light in the nanoscale is the gold-coated tapered optical fibers, widely used in Near-field Scanning Optical Microscopes (NSOMs). The optical transmission efficiency of NSOM probe tips is typically between 10^{-5} - 10^{-4} (or less). Even the performance of lasers is far away from the transmission performance required here[13].

The herein proposed technique, called APOTUS-HM[14], is a way to overcome the threshold posed by Bethe's relation. APOTUS-HM provides a transmission coefficient incomparable higher than the other established techniques, that ideally approaches unity; at the same time its basic idea is simple and quite easy to realize in practice. APOTUS-HM is founded on three main pillars : (i) unidirectionality in propagation, (ii) immunity in dispersion, and (iii) extraordinary optical transmission (EOT). The basic principle of APOTUS-HM is as follows. Using special materials it is imposed to a wave to propagate in a waveguide only forwards, unidirectionally. At the end of the waveguide there is a hole with the appropriate diameter for the focusing. When the wave arrives at the end of the waveguide, as it cannot move backwards, it is forced to pass through the hole, no matter how small the hole is, and thereby to be focused in front of it.

[13] E.g., an inexpensive 10 mW diode laser has transmission efficieny $\sim 1\%$ for a spot of $d \geq 200$ nm; however, this is still a very large diameter for what is required in the cases mentioned here.

[14] **APOTUS-HM :** Almost Perfect Optical Transmission through Unstructured Single Hole Method

Despite the simplicity of the above idea, to realize successfully an APOTUS-HM device, there are many subtle theoretical topics that must be studied and understood. In the herein thesis it is made an attempt to present how to implement efficiently the three pillars of the technique mentioned above and the basic theory behind them, giving in this way a background for the better understanding and further improving the efficiency of the technique; and of course give details of numerical simulations of basic devices (models) for the technique.

The unidirectionality and immunity of propagation in dispersion is attained using topological materials. The first and the second chapters are a general introduction to the topological materials and to some very important parameters characterizing them. In specific, the Berry phase and the Chern number are introduced and thoroughly discussed. Methods for their arithmetic computation are presented briefly. The concept of the topological material is introduced, and its relation with the time reversal symmetry and the bulk-edge correpsondence principle is discussed among others.

The third chapter is somehow a synopsis of many applications – some of them very intriguing – have the topological materials in the discipline of Photonics. APOTUS-HM could further improve the efficiency many of them.

The fourth chapter is devoted to the extraordinary optical transmission (EOT). EOT is a crucial phenomenon for APOTUS-HM since takes place extensively in it and increases the transmission coefficient. Athough EOT has been investigated quite well in many references, here it is very usefull a review of its main aspects. Also, some important results of EOT have been reproduced to gain intuition and appreciate better its role in APOTUS-HM.

In chapter five at last, the APOTUS-HM is introduced and studied, in all its aspects. Its basic ideas are presented, and models of structures that can be realized in practice are discussed. Firstly, some properties of the surface magnetoplasmons (SMPs) are briefly presented, as SMPs are another basic component of the herein technique and are not so well known as the surface plasmon polaritons (SPPs). A supporting theory, small and simple, but quite neat, concerning the temporal coupling of modes is developed[15]; this theory shows that the transmission with APOTUS-HM is in principle independent of how small the hole is (only the losses and the position of the hole play role). In some structures of APOTUS-HM it is necessary to use perfect *magnetic* conductor (PMC) as coating for the waveguide, in contrast to perfect electric conductor (PEC) as usually. PMCs do not exist in nature, they have special properties and their realization is not trivial[16]. For the integrity of the text, a section of this chapter is devoted to the basic properties of PMCs. Simulations of structures that can be used to realize APOTUS-HM devices are presented and their properties concerning transmission, propagation and dispersion are studied, in 2D and 3D cases.

In conclusive, the results are very ecnouranging and show that overcoming Bethe's threshold with APOTUS-HM is feasible, and without any serious limitation in principle. According to the developed supportive theory, the transmission through a tiny hole with APOTUS-HM does not depend on how subdiffractional the hole is, and has a transmission coefficient that can even reach unity. This result is remarkable, never observed or

[15] This small theory is in fact the theoretical foundation of how to overcome Bethe's threshold.

[16] The use of PMCs, wherever is needed, is perhaps one of the very few disadvantages of the technique. However, it is noted that in many of the examined structures the use of PMCs does not exist at all.

attained until now in the entire area of Photonics, even with the best available techniques (use of SPPs and EOT phenomena, tapered optical fibers etc) where the transmission in the deep subdiffractional regime is practically negligible[17]. Therefore, although much remains to be done for the maturity of the technique, as noted earlier APOTUS-HM could have a significant impact on the wider discipline of Photonics (EOT and topological waveguides) and its applications in micro- and nanoscale in general.

November 2023, Athens,

Konstantinos Baskourelos.

[17] For example, with the competitive techniques, transmission through $\sim \lambda_{eff}/50$ holes has transmission coefficient of order 10^{-4} or less.

3-member Examination Committee

[18]Kosmas L. Tsakmakidis, Assist. Prof. (main supervisor)
[18]Nikolaos Stefanou, Prof.
[19]Tomasz Stefański, Assoc. Prof.

7-member Examination Committee

[18]Kosmas L. Tsakmakidis, Assist. Prof.
[18]Nikolaos Stefanou, Prof.
[19]Tomasz Stefański, Assoc. Prof.,
[18]Dimosthenis Stamopoulos, Assoc. Prof.,
[18]Ioannis Lelidis, Assoc. Prof.,
[18]Dimitrios Frantzeskakis, Prof.
[20]Maria Kafesaki, Prof.

---

[18] Section of Condensed Matter Physics, Department of Physics, National and Kapodistrian University of Athens, Panepistimioupolis, GR-157 84 Athens, Greece.

[19] Faculty of Electronics, Telecommunications, and Informatics, Gdansk University of Technology, 80-233 Gdansk, Poland.

[20] Department of Materials Science and Technology, University of Crete, GR-70013 Heraklion, Crete, Greece.

1. Berry phase and Chern number : rudiments and their significance

1.1 Inroduction

The Berry phase, also referred sometimes as "geometric" or "Pancharatnam" phase, is a phase angle[1] that describes the global phase evolution of a complex vector as it moves in a path in its vector space. It was intrduced for the first time by Pancharatnam [188] in 1956 · many years later, at 1980s, Berry and others systematized and popularized the concept in a series of publications [21, 266]. A quite deep study of Berry phase can be found in texts on Differential Geometry and Topology in Physics, such as [60], [68] and [174], where other related topics (fiber bundles, connections, Berry curvatures etc.) and their relation with Berry phase are also discussed. Berry phase is used in many branches of Physics : Condensed Matter, Atomic and Molecular Physics, Nuclear Physics, Classical Optics and Photonics, to name the most prominent. In specific, its role in the theory of topological band insulators is very important as it is used extensively in the formalism of adiabatic phases which constitute a keystone in this theory.

The purpose of this chapter is to introduce in detail the Berry phase and its close related quantities Berry connection and curvature. The Chern number is also introduced; this is a crucial topological invariant which defines the open edge-state channels[2] of a topological insulator. Finally, the concept of topological properties of materials is introduced, and some fundamentals for the arithmetic computation of Berry phase and Chern number are briefly discussed. The presentation below follows [252].

1.2 Berry phase and connection, gauge invariance and parallel transport

As was mentioned above, Berry phase is an angle which indicates how a global phase accumulates as a complex vector is moving in a path[3] in its vector space. In the herein study, the interest is exclusively in phases; so, for simplicity, the complex vectors are

[1] It takes values in the interval $[0, 2\pi)$.

[2] according to the terminology of the Landauer–Büttiker formalism.

[3] The case of a closed path is that interest most, and is the most usual.

considered to be unit vectors, and identified with the ground state wavefunction of a quantum system. Typical examples is a vector representing the ground state of electrons in a molecule with fixed nuclear coordinates, a vector representing the spinor in an external magnetic field etc. In the aforementioned systems, let consider a gradual variation of the nuclear coordinates or magnetic field, in a way that the system at the end of the path returns to its starting point.

For example, consider a triatomic molecule that is almost equilateral; for some reason a distortion appears and one of the three bonds is shortened slightly and becomes the strongest; then it frees up and the distortion moves to the next bond. This process sweeps all the three bonds, one at a time, starts again and continues in this manner indefinitely. The distortion corresponds to the ground state of the molecule, and the requested is to find the phase evolution of this ground state as it moves cyclically around the bonds.

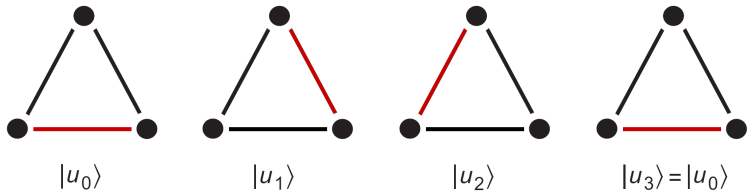


Figure 1.1: Evolution of the ground state $|u_0\rangle$ of a triatomic molecule as a distortion is carried cyclically around its bonds.

As another example, consider the evolution of the ground state of a spinor (e.g., an electron or proton) in an external magnetic field as the direction of this field varies in a closed path on the unit sphere.

Such a cyclic variation is shown schematically in Fig. 1.2, where in the path of vector are indicated $N = 8$ states, $|u_0\rangle$,...,$|u_7\rangle$, with $|u_8\rangle = |u_0\rangle$. In cases like these, the Berry phase encodes information about the phase evolution of the ground state along the path of interest.

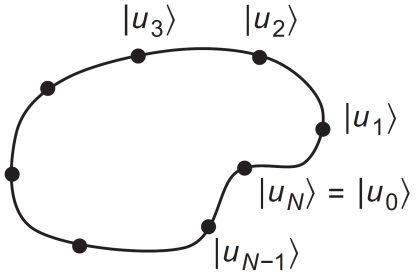


Figure 1.2: Evolution of a complex unit vector $|u\rangle$ in a closed path in parameter space.

1.2.1 Discrete case

For simplicity, let start the discussion of Berry phase with the case of discrete quantum states. Consider a system with a status vector $|u\rangle$ that varies in a closed path and its evolution in this loop is represented by N snapshot vectors, $|u_0\rangle$,...,$|u_{N-1}\rangle$. An example is the triatomic molecule sketched in Fig. 1.1 that has N=3. It is reminded that $|u_N\rangle = |u_0\rangle$ and that the vectors are complex and unit. Note also that for a complex number $z = |z_0|e^{i\phi}$, the expression Im ln $z = \phi$ discards the magnitude and gives just the phase. Then, the Berry phase ϕ is defined to be the sum of phases of the inner products of the state vectors at the successive points on the path. This can be written as

$$\phi = -\text{Im ln}\left[\,\langle u_0|u_1\rangle\langle u_1|u_2\rangle ...\langle u_{N-1}|u_0\rangle\,\right]. \tag{1.1}$$

The minus sign in (1.1) is just a convention and is not adopted universally. Note that for this definition the complex nature of the vectors is important; for real vectors the Berry phase is trivially 0 or π depending on the sign of the product; in contrast, for complex vectors can accumulate any value in [0, 2π).

As an example, consider the triatomic molecule of Fig. 1.1. Suppose that when the molecule is undistorted there are two degenarate states $|1\rangle$ and $|2\rangle$, and that the distortion breaks this degeneracy everywhere along the path; furhtermore, suppose that lower energy of the two states for the snapshots shown in Fig. 1.1 are

$$|u_0\rangle = |u_3\rangle \;= \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}, \quad |u_1\rangle = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\e^{2\pi i/3}\end{bmatrix}, \quad |u_2\rangle = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\e^{4\pi i/3}\end{bmatrix} \tag{1.2}$$

where the top and bottom elements in the column vectors are respectively the amplitudes of the basis states $|1\rangle$ and $|2\rangle$. For this configuration, the Berry phase is computed trivially to be

$$\phi = -\text{Im ln}\left[\,\langle u_0|u_1\rangle\langle u_1|u_2\rangle\langle u_2|u_0\rangle\,\right] = -\text{Im ln}\left[\,\left(\frac{e^{\pi i/2}}{2}\right)^3\,\right] = -\pi \tag{1.3}$$

or equivalently[4] $\phi = \pi$.

An important characteristic of the Berry phase is that it is independent of the choices made for the phases of the individual vectors $|u_j\rangle$. Specifically, if a new set of N states $|\tilde{u}_j\rangle$ is introduced in the form

$$|\tilde{u}_j\rangle = e^{-i\beta_j}|u_j\rangle \tag{1.4}$$

where β_j is real, then the Berry phase remains the same as for the old set $|u_j\rangle$. In (1.4) the old vectors are related to the new ones by a j-dependent phase rotation β_j. This operation is a *gauge tranformation* in the Berry phase theory[5]; the Berry phase remains

[4] As it is explained below, the Berry phase is well defined only modulo 2π.

[5] The name refers to the use of the same term in the theory of Electromagnetism. In all cases, a particular choice of gauge may influence the intermediate results of a calculation but does not affect any physically meaningful prediction.

the same since any vector used in (1.1) appears once in a ket and once in a bra, so that the phases $e^{\pm\beta_j}$ cancel out. The *gauge invariance* of the Berry phase insinuates strongly that it may be related with some physically observable phenomena.

A detail not mentioned in the above discussion is the need to impose a branch choice on the definition of $\mathrm{Im}\ln z$, as by restricting it to the interval $-\pi < \phi \leq \pi$. Adopting this convention, (1.1) always gives a Berry phase lying in this interval, while the ostensibly equivalent expression[6]

$$\phi = -\sum_{j=0}^{N-1} \mathrm{Im}\ln\langle u_j|u_{j+1}\rangle \tag{1.5}$$

can yield a result that differs by an integer multiple of 2π. From the viewpoint that ϕ is just a shorthand for a phase angle, only $\cos\phi$ and $\sin\phi$ matter, and this distinction can be safely ignored. However, in all practical calculations the phase angles are normally mapped onto some interval on the real axis, and can be claimed only that the Berry phase should be gauge-invariant modulo 2π in the context of an expression like that of (1.5). This subject will be examined extensivelly bellow.

As can be seen from (1.1) and (1.5), the information carried by the Berry phase is pumped only from the phase of the involved vectors. However, in the same manner a corresponding function $-\mathrm{Re}\ln\prod_{j=0}^{N-1}\langle u_j|u_{j+1}\rangle$ can be defined for their magnitude[7]. This function measures how much the nature of the states varies from point to point along the path; in contrast, the Berry phase is related only to the relative phases along the path.

Another remarkable characteristic of the Berry phase is its relation with the parallel transport of the state vectors in a closed path; in fact, it could alternatively be defined in this context. Consider a set of states $|u_0\rangle$, $|u_1\rangle$,...,$|u_N\rangle$ without a specific phase relation between them. Using a concept from Differential Geometry, a new set of "parallel transported" states $|\bar{u}_0\rangle$, $|\bar{u}_1\rangle$,...,$|\bar{u}_N\rangle$ can be defined as follows. Set $|\bar{u}_0\rangle = |u_0\rangle$. Then set $|\bar{u}_1\rangle$ to be $|u_1\rangle$ times a phase chosen such that $\langle\bar{u}_0|\bar{u}_1\rangle$ is real and positive. Similarly, set $|\bar{u}_2\rangle$ such that $\langle\bar{u}_1|\bar{u}_2\rangle$ is real and positive. Continue in this manner for all the vectors of the path, imposing the constraint

$$\mathrm{Im}\ln\langle\bar{u}_j|\bar{u}_{j+1}\rangle = 0 \tag{1.6}$$

for the successive vectors. At the end, set $|\bar{u}_N\rangle$ such that $\langle\bar{u}_{N-1}|\bar{u}_N\rangle$ is real and positive. This process is a *parallel transport gauge (PT gauge) transformation* in the sense of Berry phase[8].

Assume that a set of states[9] is given, forming a closed path as in Fig. 1.2. The

---

[6] without a specific choice for the branch of ln.

[7] It is reminded that the vectors are unit, so their inner products are smaller or maximally equal to unity.

[8] As already noted, the term "parallel transport" comes from Differential Geometry, where it is implemented choosing a local orthonormal basis of vectors at each point along a path on a curved manifold, in such a way that the basis is "as aligned as possible" with its neighboring vectors everywhere along the path. Here, the request "as aligned as possible" is interpreted in the sense of phase equality. Evidently, (1.6) makes the relative phase of the two vectors to be 0 or π making them "parallel".

[9] From now on "states" and "vectors" will mean "state vectors" and will be used alternatively as an abbreviation.

two vetors $|u_0\rangle$ and $|u_N\rangle$ are identical. Let this set is subjected to a PT gauge as described above. Then, the two vectors $|\bar{u}_0\rangle$ and $|\bar{u}_N\rangle$ correspond to the same physical state but they generally differ by a phase. This phase difference between $|\bar{u}_0\rangle$ and $|\bar{u}_N\rangle$ is just the Berry phase. It is easy to prove this. Recall that (1.1) is gauge-invariant; then, instead of the initial vectors, the PT gauge vectors $|\bar{u}_0\rangle$,...,$|\bar{u}_{N-1}\rangle$ can be used, that is $\phi = -\mathrm{Im}\ln\big[\,\langle\bar{u}_0|u_1\rangle...\langle\bar{u}_{N-1}|\bar{u}_0\rangle\,\big]$. Since $|\bar{u}_0\rangle$ and $|\bar{u}_N\rangle$ differ only by a phase, the $|\bar{u}_0\rangle$ at the end of the product can be replaced[10] by $|\bar{u}_N\rangle\langle\bar{u}_N|\bar{u}_0\rangle$ to get $\phi = -\mathrm{Im}\ln\big[\,\langle\bar{u}_0|u_1\rangle...\langle\bar{u}_{N-1}|\bar{u}_N\rangle\langle\bar{u}_N|\bar{u}_0\rangle\,\big]$. Then all inner products are real and positive[11] except the last, so

$$\phi = -\mathrm{Im}\ln\langle\bar{u}_N|\bar{u}_0\rangle\,. \tag{1.7}$$

For example, for the set of states (1.2) it is

$$|\bar{u}_0\rangle = \begin{bmatrix}1\\1\end{bmatrix},\ \ |\bar{u}_1\rangle = \begin{bmatrix}e^{-\pi i/3}\\ e^{\pi i/3}\end{bmatrix},\ \ |\bar{u}_2\rangle = \begin{bmatrix}e^{-2\pi i/3}\\ e^{2\pi i/3}\end{bmatrix},\ \ |\bar{u}_3\rangle = \begin{bmatrix}-1\\-1\end{bmatrix} \tag{1.8}$$

where the irrelevant normalization coefficients have been dropped.
The Berry phase is $\phi = -\mathrm{Im}\ln\big[\,\langle\bar{u}_N|\bar{u}_0\rangle\,\big] = \pi$ as before.

It is point out that the PT gauge is not unique, since the phase of the initial vector $|\bar{u}_0\rangle$ can be choosen arbitrarily. The choice of initial phase propagates into $|\bar{u}_N\rangle$ through (1.6); however, it does not affect the value of ϕ resulting from (1.7).

For the closed paths studying here, the PT gauge has an annoying disadvantage : it produces a discontinuity on the vectors at the end of the path where the end and the starting point rejoined. This discontinuity can be smoothed by constructing a *twisted parallel transport (TPT) gauge* by starting from the PT gauge and applying phase twists[12]

$$|\tilde{u}_j\rangle = e^{-ij\phi/N}|\bar{u}_j\rangle\,. \tag{1.9}$$

This gauge no longer produces discontinuity at the end of the loop. It has the property that $\mathrm{Im}\ln\langle\tilde{u}_j|\tilde{u}_{j+1}\rangle$ has the value $-\phi/N$ at every point on the loop, in consistency with (1.5). In fact, what it does is to distribute uniformly the phase evolution along the loop in such a way as to smooth the gauge discontinuity that otherwise would occur at the end of the loop.

The TPT gauge seems to be quite restricted; however, it is less restricted than the simple PT gauge for the following reason. In addition to rotating the phase of the initial state $|\bar{u}_0\rangle$ (which results to a global rotation of all phases), now there is the possibility to replace ϕ by $\phi + 2\pi m$ (where m is an integer) in (1.9). For example, taking $m = 1$ changes all the $\mathrm{Im}\ln\langle\tilde{u}_j|\tilde{u}_{j+1}\rangle$ by $-2\pi/N$, which for large N is much less than 2π. This means that there is freedom to choose different ways to smooth the phase discontinuity

[10] If a vector $|w\rangle$ is unit as it happens here, then for a product $\langle a|b\rangle$ holds $\langle a|b\rangle = \langle a|w\rangle\langle w|b\rangle$ because multiplying with $|w\rangle$ does not change the magnitude of the product, and also the phase does not affected since $|w\rangle$ appears both as a bra and a ket and the phase changes cancel out.

[11] Because of the PT gauge, the pairs of new vectors have zero phase differences.

[12] The TPT gauge distributes the additional phase ϕ on the vectors of the PT gauge, not to the initial vectors. That is, first the PT gauge is applied, and then the twisted one.

such that $\mathrm{Im}\ln\langle\tilde{u}_j|\tilde{u}_{j+1}\rangle$ is identical for each pair of successive vectors; each such way is a different but equivalent TPT gauge. Usually the TPT gauge that makes $\mathrm{Im}\ln\langle\tilde{u}_j|\tilde{u}_{j+1}\rangle$ minimum is choosen, but this is only a convenient choice, not a fundamental restriction. The gauge choice for the vectors (1.2) is also an example of a TPT gauge.

1.2.2 Continuous case and Berry connection

The discrete formulation of Berry phase presented above can reasonably be extended to the continuous case. The process is shown schematically in Fig. 1.3. The continuous limit is obtained by increasing unlimitedly the number of points corresponding to states along a path. The path is parametrized by a real variable $\lambda \in [0, 1]$, and for the most usual case, a closed path, it is $|u_{\lambda=0}\rangle = |u_{\lambda=1}\rangle$. A state $|u_\lambda\rangle$ is considered to be a smooth (and hence differentiable as much as desired) function of λ.

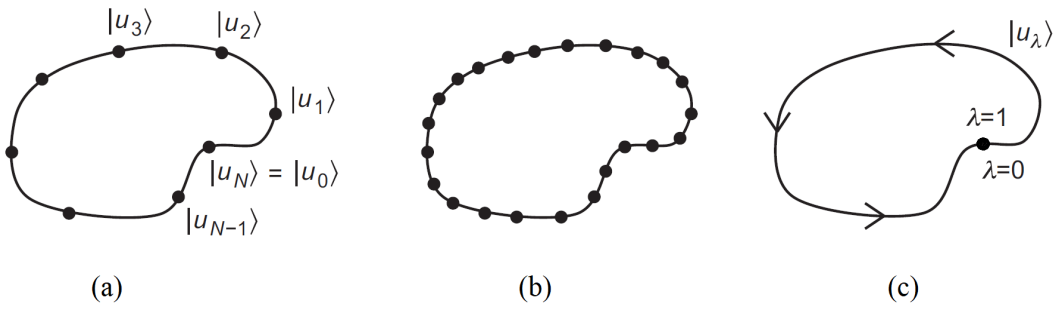


Figure 1.3: (a) Evolution of a state $|u_0\rangle$ in N discrete steps on a closed path. (b) Increasing the number of states on the path to reach the continuous limit. (c) Continuum limit, where a parameter λ varies in [0, 1] and $|u_{\lambda=0}\rangle = |u_{\lambda=1}\rangle$.

The derivation of the continuous expression of the Berry phase starts with (1.5). Using Taylor series for $|u_{\lambda+d\lambda}\rangle$ and for the ln, it is

$$\begin{aligned}\ln\langle u_\lambda|u_{\lambda+d\lambda}\rangle &= \ln\langle u_\lambda|\left(|u_\lambda\rangle + d\lambda\,\frac{d|u_\lambda\rangle}{d\lambda} + ...\right)\\ &= \ln\left(1 + d\lambda\,\langle u_\lambda|\partial_\lambda u_\lambda\rangle + ...\right)\\ &= d\lambda\,\langle u_\lambda|\partial_\lambda u_\lambda\rangle + ...\end{aligned} \tag{1.10}$$

where the relation $\ln(1+x) \simeq x$ was used[13], and $\langle u_\lambda|u_\lambda\rangle = 1$ since the vectors are unit. ∂_λ is simply a shorthand for $d/d\lambda$ and "..." indicates terms of second order and higher in $d\lambda$. Taking the continuum limit of (1.5) and using the above, these terms are discarded, and is obtained

[13] It holds for $-1 < x \leq 1$.

$$\phi = -\mathrm{Im} \oint \langle u_\lambda | \partial_\lambda u_\lambda \rangle \, d\lambda \,. \tag{1.11}$$

But $\langle u_\lambda | \partial_\lambda u_\lambda \rangle$ is purely imaginary because

$$2\mathrm{Re}\langle u_\lambda | \partial_\lambda u_\lambda \rangle = \langle u_\lambda | \partial_\lambda u_\lambda \rangle + \langle \partial_\lambda u_\lambda | u_\lambda \rangle = \partial_\lambda \langle u_\lambda | u_\lambda \rangle = 0 \,, \tag{1.12}$$

hence (1.11) can be written as[14]

$$\phi = \oint \langle u_\lambda | i \partial_\lambda u_\lambda \rangle \, d\lambda \,. \tag{1.13}$$

This is the expression for the Berry phase in the continuous case, as introduced in [21] and [266]. The integrand of (1.13) is the so called *Berry connection*, also known as *Berry potential*[15]

$$A(\lambda) = \langle u_\lambda | i \partial_\lambda u_\lambda \rangle = -\mathrm{Im} \langle u_\lambda | \partial_\lambda u_\lambda \rangle \,, \tag{1.14}$$

with which the Berry phase is written as

$$\phi = \oint A(\lambda) \, d\lambda \,. \tag{1.15}$$

Below, it will be examined how the Berry phase and connection behave under a gauge tranformation. In analogy with (1.4), a gauge in the continuous case takes the form

$$|\tilde{u}_\lambda \rangle = e^{-i\beta(\lambda)} |u_\lambda \rangle \tag{1.16}$$

where $\beta(\lambda)$ is a real, continuous function, and $\beta'(\lambda) = d\beta / d\lambda$.
It is find easily that

$$\tilde{A}(\lambda) = \langle \tilde{u}_\lambda | i \partial_\lambda \tilde{u}_\lambda \rangle = \langle u_\lambda | e^{i\beta(\lambda)} \, i \partial_\lambda \, e^{-i\beta(\lambda)} | u_\lambda \rangle = \langle u_\lambda | i \partial_\lambda u_\lambda \rangle + \beta'(\lambda) \,. \tag{1.17}$$

This means that Berry connection is transformed under a gauge change according to the rule

$$\tilde{A}(\lambda) = A(\lambda) + \beta'(\lambda) \tag{1.18}$$

and thus it is not gauge-invariant.

[14] When a complex z is purely imaginary, it holds $\mathrm{Im}(z) = -iz$.

[15] The term "connection" stems from Differential Geometry, while "potential" indicates an analogy with the vector potential in Electromagnetism. In the Berry phase sense the names are used interchangeably.

Concerning the Berry phase, note that for a closed path must holds $|\tilde{u}_{\lambda=1}\rangle = |\tilde{u}_{\lambda=0}\rangle$, just as it is for $|u_\lambda\rangle$. But then (1.16) implies that

$$\beta(1) = \beta(0) + 2\pi k \tag{1.19}$$

where k is an integer. Then

$$\int_0^1 \beta'(\lambda)\, d\lambda = \beta(\lambda = 1) - \beta(\lambda = 0) = 2\pi k\,. \tag{1.20}$$

Thus, replacing $A(\lambda)$ by $\tilde{A}(\lambda)$ in (1.15), and using (1.18), gives

$$\tilde{\phi} = \phi + 2\pi k\,. \tag{1.21}$$

This means that the Berry phase ϕ is gauge-invariant modulo 2π, i.e., it is gauge-invariant when considered as a phase angle.

As in the discrete case, the Berry phase can be regarded to be the residue of phase that remains after a parallel transport around a closed path. In the continuous case, and in correspondence with (1.6), a PT gauge is one in which the Berry connection $A(\lambda)$ vanishes :

$$\tilde{A}(\lambda) = \langle \tilde{u}_\lambda | i\partial_\lambda \tilde{u}_\lambda \rangle = 0\,. \tag{1.22}$$

Under such a gauge, the Berry phase is just the phase difference at the end of the closed path,

$$\phi = -\mathrm{Im}\ln\langle \bar{u}_{\lambda=1} | \bar{u}_{\lambda=0} \rangle\,, \tag{1.23}$$

exactly as in the discrete case, (1.7).
A TPT gauge can also be imposed, $|\tilde{u}_\lambda\rangle = e^{-i\phi\lambda}|\tilde{u}_\lambda\rangle$, in correspondence with (1.9) ˙ this results the $\tilde{A}(\lambda)$ to be constant on the closed path.

Although not immediately evident, the property that Berry phase is gauge-invariant modulo 2π is important. The quantum probabilities are proportional to the norm squared of an amplitude, giving thus a tendency to think that the phase is indifferent. But this is not true ˙ the phases can lead to interference phenomena that are physically important. For example, let consider two identical copies of a system are prepared, subjected to parallel transport along different paths in a parameter space, and then recombined ˙ then the resulting phase difference can lead to physical and measurable interference effects.

The Berry phase is mainly of interest for the evolution of states along closed paths ˙ however, it has sense and applications for open paths too. Let consider an open path, like that sketched in Fig. 1.4a. For such an open path, the Berry phase is defined as

$$\phi = \int_{\lambda_i}^{\lambda_f} A(\lambda)\, d\lambda \tag{1.24}$$

where λ is a scalar running from λ_i to λ_f determining the evolution along a path in a higher-dimensional parameter space, labeled as (λ_x , λ_y) in the figure.

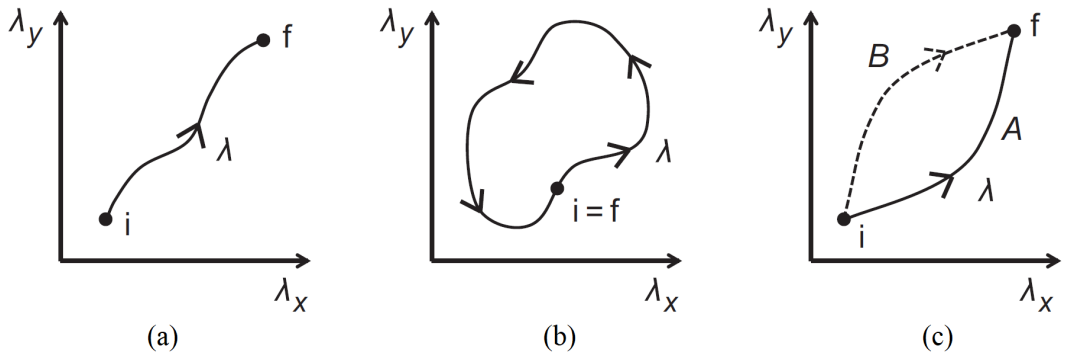


Figure 1.4: (a) An open path connecting initial point 'i' to final point 'f' in a 2D parameter space. (b) A closed path. (c) A pair of open paths A and B with common initial and final points, such that $A-B$ (i.e., traverse A, then traverse the reverse of B) is a closed path.

The Berry phase for an open path is not gauge-inavariance˙ a gauge like (1.16) causes a change $\Delta\phi = \beta_f - \beta_i$. The Berry phase is gauge-inavariance modulo 2π only when the path is closed, as in Fig. 1.4b. However, there is an interesting case, sketched in Fig. 1.4c. When a system is moving from λ_i to λ_f along two different paths A and B, the change[16] in Berry phase $\Delta\phi = \phi_B - \phi_A$ is also gauge-invariant. This results trivially : traversing first path B changes the phase by ϕ_B, then path A in the reverse direction changes the phase by ϕ_A, a total change $\Delta\phi = \phi_B - \phi_A$. But this is equivalent to circulate around a closed path, for which $\Delta\phi$ is gauge-invariant.

Concerning again closed paths, it is pointed out that the integer k which appears in (1.19) can be used to classify topologically all possible gauge transformations of type (1.16). k is a *winding number*˙ it indicates how many times $e^{-i\beta}$ circulates around the unit circle in the complex plane as λ circulates on the path.

The gauge changes with $k = 0$, Fig. 1.5a, are called *progressive* gauge transformations˙ they have the property that the gauge function $\beta(\lambda)$ can be deformed homotopically to the identity transformation[17]. In contrary, the gauge changes with $k \neq 0$ are called $radical$, Fig. 1.5c,d.

It is emphasized that the Berry phase itself is not a quantized quantity and cannot be used as a topological index. In contrary, the winding number is quantized and can serve as a topological index for the set of gauge transformations on a closed path.

1.2.3 A simple example of Berry phase computation

As mentioned earlier, a common application of Berry phase is to count the phase that accumulates during the evolution of the ground state wavefunction of a quantum system. In such a case, the ground state $|u_\lambda\rangle$ and the Hamiltonian $\widehat{H}_\lambda$ of the system are

[16] i.e., the relative phase between the two points.

[17] i.e., $\beta = 0$ independent of λ.

parametrized by λ. A typical example is the ground state of electrons in a molecule, where λ describes the variation of a coordinate or the component of an external field. The variation of λ must be such that $|u_\lambda\rangle$ and H_λ evolve with time in a continuous way. In specific, it is adopted the condition that the variation is slow enough that the state vector is approximated reliably by the static solution $|u_\lambda\rangle$ at the current value of λ, in all the path. This is the so called *adiabatic approximation*, and *adiabatic evolution*. Essentially, adiabatic approximation means that the variation is continuous.

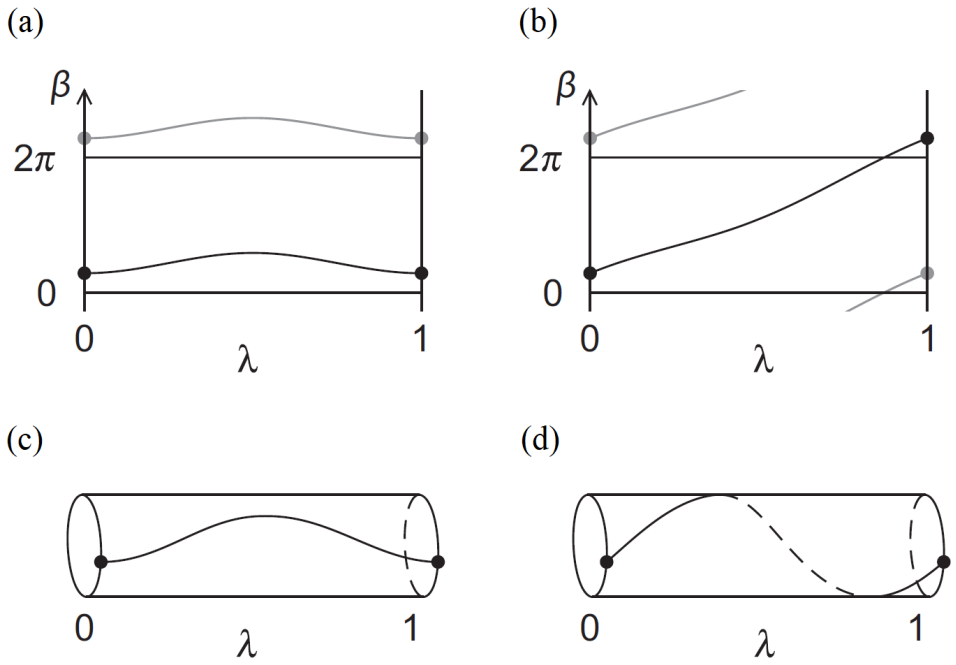


Figure 1.5: (reprinted from [252]). Possible behaviors of the function $\beta(\lambda)$ defining a gauge transformation via (1.16). (a) "Progressive" gauge tranformation. At the end of the loop β returns to itself. (b) "Radical" gauge tranformation. At the end of the loop β is shifted by a multiple of 2π. Gray lines indicate 2π-shifted periodic loops. (c), (d) Same as (a) and (b) but plotted on the surface of a cylinder to signify better the non-zero winding of the radical gauge transformation in (b).

An illustrative, simple example is a spin-$\frac{1}{2}$ particle (e.g., an electron), at rest, under a uniform magnetic field $\mathbf{B} = B\hat{\mathbf{n}}$ along the $\hat{\mathbf{n}}$ direction. The Hamiltonian of this particle is [178]

$$\widehat{H} = -\boldsymbol{\mu} \cdot \mathbf{B} = \mu_B B\, \boldsymbol{\sigma} \cdot \hat{\mathbf{n}} \tag{1.25}$$

where $\boldsymbol{\mu}$ is the magnetic moment, μ_B is the Bohr magneton and $\boldsymbol{\sigma} = (\sigma_x\,, \sigma_y\,, \sigma_z)$ are the Pauli matrices. The ground state $|u_{\mathbf{B}}\rangle$ is an eigenstate of $\boldsymbol{\sigma} \cdot \hat{\mathbf{n}}$, with its spin along $\hat{\mathbf{n}}$ direction ˙ therefore, $|u_{\mathbf{B}}\rangle$ is dependent only on the direction $\hat{\mathbf{n}}$ of $\mathbf{B}$, not on the magnitude.

This is emphasized writting $|u_{\hat{\mathbf{n}}}\rangle$ instead of $|u_{\mathbf{B}}\rangle$ from now on. The requested is to find the Berry phase of $|u_{\hat{\mathbf{n}}}\rangle$ as $\hat{\mathbf{n}}$ moves on a closed path of a spherical surface, the so called *Bloch sphere*[18] [12].

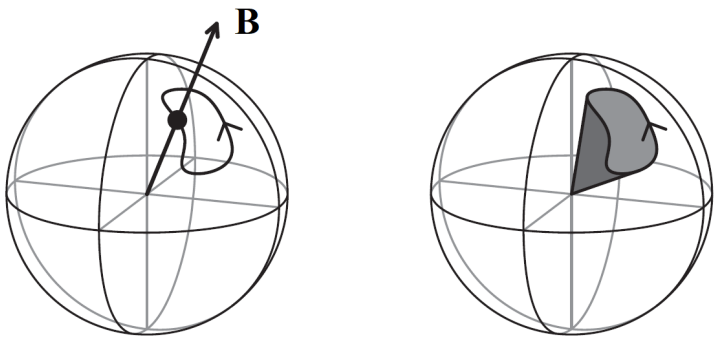


Figure 1.6: (reprinted from [252]). (a) Evolution of an applied magnetic field **B** along a closed path on a spherical surface (Bloch sphere). (b) Solid angle traversed during the moving of **B** in (a).

The answer to this for the general case will be given later. For now, the case of an octant of the sphere will be examined. Let $\hat{\mathbf{n}}$ initially is directed along $\hat{\mathbf{z}}$, then it is rotated successively to $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ and at last to $\hat{\mathbf{z}}$ again, forming a closed path. The Berry phase for this path is given via (1.1) as

$$\phi = -\mathrm{Im}\ln\left[\langle\uparrow_{\hat{\mathbf{z}}}|\uparrow_{\hat{\mathbf{x}}}\rangle\langle\uparrow_{\hat{\mathbf{x}}}|\uparrow_{\hat{\mathbf{y}}}\rangle\langle\uparrow_{\hat{\mathbf{y}}}|\uparrow_{\hat{\mathbf{z}}}\rangle\right] \tag{1.26}$$

where $|\uparrow_{\hat{\mathbf{n}}}\rangle$ is the spinor, with spin in $\hat{\mathbf{n}}$ direction, and likewise in $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$.
Such a spinor has the form [12]

$$|\uparrow_{\hat{\mathbf{n}}}\rangle = \begin{bmatrix} \cos(\theta/2) \\ \sin(\theta/2)e^{i\phi} \end{bmatrix} \tag{1.27}$$

where θ is the polar and ϕ the azimuthal angle of $\hat{\mathbf{n}}$. Therefore, the states of the spinor in (1.26) are

$$|\uparrow_{\hat{\mathbf{x}}}\rangle = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}, \quad |\uparrow_{\hat{\mathbf{y}}}\rangle = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\i\end{bmatrix}, \quad |\uparrow_{\hat{\mathbf{z}}}\rangle = \begin{bmatrix}1\\0\end{bmatrix}. \tag{1.28}$$

For the Berry phase the normalization factors are indifferent and can be ommitted ; thus, (1.26) gives $\phi = -\mathrm{Im}\ln\left[(1)(1+i)(1)\right] = -\pi/4$.

[18] Practically, this is the unit sphere in spin space.

This result concerns the octant of the spherical surface and happens to be exact ˙ it does not change if more intermediate states[19] are used. To compute the Berry phase for a general geometric region on the spherical surface, the concept of Berry curvature must be introduced ˙ this is done in next section.

1.3 Berry curvature and Chern number

1.3.1 Berry curvature and Berry flux

The concept of Berry connection can easily be generalized in a multi-parameter space ˙ for convenience a two-parameter space will be examined but the generalization to more is straightforward. Let consider such a space, sketched in Fig. 1.7. The parameter of state vector $|u_{\boldsymbol{\lambda}}\rangle$ is a vector itself, $\boldsymbol{\lambda} = (\lambda_x, \lambda_y)$, and the definiton (1.14) for Berry connection becomes

$$\mathbf{A} = \langle u_{\boldsymbol{\lambda}} | i \nabla_{\boldsymbol{\lambda}} u_{\boldsymbol{\lambda}} \rangle \,, \tag{1.29}$$

where $\mathbf{A} = (A_x, A_y)$, or in componentwise form :

$$A_\mu = \langle u_{\boldsymbol{\lambda}} | i \partial_\mu u_{\boldsymbol{\lambda}} \rangle \tag{1.30}$$

where $\partial_\mu = \partial / \partial_\lambda$ and $\mu = x, y$. Now, the Berry phase in the definition (1.15) is written as line integral long the path L, as

$$\phi = \oint_L \mathbf{A} \cdot d\boldsymbol{\lambda} \,. \tag{1.31}$$

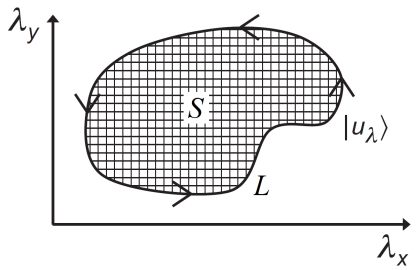


Figure 1.7: State vector in a two-parameter space. The parameter of state vector is a vector itself, $\boldsymbol{\lambda} = (\lambda_x, \lambda_y)$.

In (1.31) the Berry phase concerns the region S which is bounded by the path L. The *Berry curvature* $\Omega(\boldsymbol{\lambda})$ is simply the Berry phase per unit area in space (λ_x, λ_y). If the region S is discretized in a mesh, as sketched in Fig. 1.7, the Berry curvature is the

[19] i.e., a more dense discretization of the path.

Berry phase around a cell divided by the area of that cell. In the continuum limit, the Berry curvature becomes the curl of the Berry connection, that is

$$\mathbf{\Omega}(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}} \times \mathbf{A}(\boldsymbol{\lambda}) \,. \tag{1.32}$$

In a two-parameter space, noting that $\langle \partial_y u | \partial_x u \rangle^* = \langle \partial_x u | \partial_y u \rangle$ and cancelling terms of the form $\langle u | \partial_x \partial_y u \rangle$, this is written as

$$\Omega_{xy}(\boldsymbol{\lambda}) = \partial_x A_y - \partial_y A_x = -2\mathrm{Im}\langle \partial_x u | \partial_y u \rangle \,. \tag{1.33}$$

Since Ω is defined as curl, the Stokes' theorem is applicable and a quantity reasonable called *Berry flux* Φ_S can be defined through the surface S. In specific, it is[20]

$$\begin{aligned} \Phi_S &= \int_S \mathbf{\Omega}(\boldsymbol{\lambda}) \cdot d\mathbf{S} \\ &= \oint_L \mathbf{A} \cdot d\boldsymbol{\lambda} = \phi_L \,. \end{aligned} \tag{1.34}$$

This means that the Berry flux through the surface equals the Berry phase around its boundary. When the surface is discretized in a mesh, Stokes' theorem states that summing up the circulation of the Berry connection $\mathbf{A}$ along all cells constituting the surface, this will give the Berry phase around the boundary of the surface[21].

Since the Berry curvature is the curl of $\mathbf{A}$, it is almost evident that is gauge-invariant. A gauge change in 2D parameter space is of the form $|\tilde{u}_{\boldsymbol{\lambda}}\rangle = e^{-i\beta(\boldsymbol{\lambda})}|u_{\boldsymbol{\lambda}}\rangle$ and (1.18) is generalized to

$$\tilde{\mathbf{A}}(\boldsymbol{\lambda}) = A(\boldsymbol{\lambda}) + \nabla_{\boldsymbol{\lambda}}\, \beta(\boldsymbol{\lambda}) \,. \tag{1.35}$$

But then, $\nabla_{\boldsymbol{\lambda}} \times \mathbf{A}(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}} \times \tilde{\mathbf{A}}(\boldsymbol{\lambda})$ and from (1.32) stems than Ω remains the same[22].

The definition of Berry curvature in a space with more than two parameters[23] is straightforward. In this case, the parameter $\boldsymbol{\lambda}$ and Berry connection $\mathbf{A}$ are n-component vectors, that is $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)$, and $\mathbf{A} = (A_1, ..., A_n)$. The Berry curvature from (1.33) becomes

$$\Omega_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu = -2\mathrm{Im}\langle \partial_\mu u | \partial_\nu u \rangle \tag{1.36}$$

and is a second order antisymmetric tensor[24].

---

[20] The boundary L is traversed in the positive sense of circulation.

[21] The contribution from any internal side of the cells vanishes because the contributions from every two neighboring cells are cancelled out.

[22] It is pointed out the remarkable analogy in gauge-invariance properties between the Berry curvature and connection, versus the magnetic field and vector potential. From this analogy stems the alternative name "Berry potential" for the Berry connection.

[23] i.e., more than two dimenions.

[24] Note that since Ω is antisymmetric tensor, the order of indices μ and ν matters : changing the order, changes the sign of Ω.

In this more than two parameters space, the Stokes' theorem holds for any surface S of a 2D submanifold, specifically

$$\phi = \oint_L \mathbf{A} \cdot d\boldsymbol{\lambda} = \int_S \Omega_{\mu\nu}\, dS_\mu \wedge dS_\nu \tag{1.37}$$

where $dS_\mu \wedge dS_\nu$ is the area element of the surface on submanifold and P is the boundary of S. For this formalism, a 3D parameter space is quite supervisory and familliar. In this case Berry curvature can be written as $\mathbf{\Omega} = -\mathrm{Im}\langle\nabla_{\boldsymbol{\lambda}} u| \times |\nabla_{\boldsymbol{\lambda}} u\rangle$ and considered to be a vector ˙ then Stokes' theorem takes its known familiar form.

Besides the above formal formulas for the Berry curvature, Berry himself provided in [21] a formula very convenient for practical use in quantum systems. To derive this in a 3D parameter space, (1.33) is written for an eigenstate $|n\rangle$ as

$$\Omega_j = -\mathrm{Im}\, \epsilon_{jkl} \langle \partial_k n | \partial_l n \rangle \,. \tag{1.38}$$

where ϵ_{jkl} is the alternating symbol. Using the completeness of the eigenstates of the Hamiltonian, $\sum_i |n_i\rangle\langle n_i| = \mathbb{1}$, the unity operator $\mathbb{1}$ is inserted in (1.38), giving thus

$$\mathbf{\Omega}^{(n)} = -\mathrm{Im} \sum_{n' \neq n} \langle \nabla n | n' \rangle \times \langle n' | \nabla n \rangle \tag{1.39}$$

where the superscript (n) indicates the eigenstate of interested $|n\rangle$ that is evolved, and the subscript $\boldsymbol{\lambda}$ indicating the parameters is skipped for simplicity in the notation. The term $n' = n$ in the sum is omitted because it is zero since the conservation of norm implies $\langle \nabla n | n \rangle = -\langle n | \nabla n \rangle$. To calculate $\langle n' | \nabla n \rangle$, start from the definition of the eigenstate $|n\rangle$, act on both sides with ∇, and then project onto $\langle n'|$. Successively it is :

$$\begin{aligned} \widehat{H}\,|n\rangle &= E_n\,|n\rangle \\ (\nabla \widehat{H})|n\rangle + \widehat{H}\,|\nabla n\rangle &= \underbrace{(\nabla E_n)|n\rangle}_{=0} + E_n |\nabla n\rangle \\ \langle n'|\nabla \widehat{H}|n\rangle + \langle n'|\widehat{H}|\nabla n\rangle &= E_n \langle n'|\nabla n\rangle \end{aligned} \tag{1.40}$$

where $\widehat{H}$ is the Hamiltonian and E_n the eigenvalues of the system.
Furthermore, act with $\widehat{H}$ on the left of (1.40), rearrange and substitute into (1.39) ˙ the result finally is

$$\mathbf{\Omega}^{(n)} = -\mathrm{Im} \sum_{n' \neq n} \frac{\langle n|\nabla \widehat{H}|n'\rangle \times \langle n'|\nabla \widehat{H}|n\rangle}{(E_n - E_{n'})^2} \,. \tag{1.41}$$

This relation gives the Berry curvature of a quantum system being in eigenstate $|n\rangle$ as a function of the rest eigenstates. It is evident that $\mathbf{\Omega}^{(n)}$ is gauge-invariant, as expected. Also, as can be seen from (1.41), if $\mathbf{\Omega}^{(n)}$ has monopole sources, these are points of degeneracy.

An implication of (1.41), is that the sum of the Berry curvatures of all eigenstates of a Hamiltonian is zero. Specifically, let a Hamiltonian $\widehat{H}$ which is discrete along a closed path in the parameters space. Adding all the phases of the eigenstates gives

$$\begin{aligned}\sum_n \mathbf{\Omega}^{(n)} &= -\mathrm{Im}\sum_n \sum_{n'\neq n} \frac{\langle n|\nabla\widehat{H}|n'\rangle \times \langle n'|\nabla\widehat{H}|n\rangle}{(E_n - E_{n'})^2} \\ &= -\mathrm{Im}\sum_n \sum_{n'<n} \frac{1}{(E_n - E_{n'})^2}\Big[\langle n|\nabla\widehat{H}|n'\rangle \times \langle n'|\nabla\widehat{H}|n\rangle \\ &\qquad + \langle n'|\nabla\widehat{H}|n\rangle \times \langle n|\nabla\widehat{H}|n'\rangle\Big] = \mathbf{0}\,. \end{aligned} \tag{1.42}$$

The last equality stems from the antisymmetry of the cross product of any two vectors, $\mathbf{a}\times\mathbf{b} = -\mathbf{b}\times\mathbf{a}$.

1.3.2 A simple example of Berry curvature computation

The example discussed in §1.2.3 to demonstrate a computation of Berry phase, can also be used to demonstrate a computation of Berry curvature and flux.
Let consider again the spinor

$$|\uparrow_{\hat{\mathbf{n}}}\rangle = \begin{bmatrix} \cos(\theta/2) \\ \sin(\theta/2)e^{i\phi} \end{bmatrix} \tag{1.43}$$

in the same scenery as in §1.2.3 (uniform magnetic field etc). The requested is to find the Berry flux of $|u_{\hat{\mathbf{n}}}\rangle$ after $\hat{\mathbf{n}}$ has completed a full (closed) path on an octant of the Bloch sphere.

In representation (1.43) there is a gauge implicitly selected˙ this gauge makes $|\uparrow_{\hat{\mathbf{n}}}\rangle$ continuous and smooth in the north pole ($\theta = 0$) of the Bloch sphere[25]. Very close to $\theta = 0$, the dependence of $|\uparrow_{\hat{\mathbf{n}}}\rangle$ on λ can be considered to be $\boldsymbol{\lambda} = (n_x, n_y)$,
with $\hat{\mathbf{n}} = (n_x, n_y, \sqrt{1-n_x^2-n_y^2})$ and $|\uparrow_{\hat{\mathbf{n}}}\rangle$ is approximately[26]

$$|\uparrow_{\hat{\mathbf{n}}}\rangle \simeq \begin{bmatrix} 1 \\ (n_x + in_y)/2 \end{bmatrix}, \quad |\partial_{n_x}\uparrow_{\hat{\mathbf{n}}}\rangle = \frac{1}{2}\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad |\partial_{n_y}\uparrow_{\hat{\mathbf{n}}}\rangle = \frac{1}{2}\begin{bmatrix} 0 \\ i \end{bmatrix}. \tag{1.44}$$

Setting the above in (1.33) gives $\Omega = -\frac{1}{2}$, at $\theta = 0$ (i.e., at $\hat{\mathbf{n}} = \hat{\mathbf{z}}$).

Furthermore, in free space the properties of a spinor are intrinsically isotropic˙ this permits to evaluate Ω in any direction $\hat{\mathbf{n}}$ in the above way, using a coordinate system $O\hat{\mathbf{x}}'\hat{\mathbf{y}}'\hat{\mathbf{z}}'$ where $\hat{\mathbf{z}}'$ is aligned to $\hat{\mathbf{n}}$. Therefore, for the herein physical system and the unit sphere, it holds $\Omega = -\frac{1}{2}$ in any direction.

---

[25] This does not mean that $|\uparrow_{\hat{\mathbf{n}}}\rangle$ is smooth everywhere in the sphere. In fact, this gauge choice introduces a singularity in the south pole ($\theta = \pi$), where $|\uparrow_{\hat{\mathbf{n}}}\rangle$ has singular dependence on ϕ. This issue will be examined in detail later.

[26] first order Taylor expansion.

Having computed the Berry curvature, it is an easy task to compute the Berry phase for a path along an octant of the unit sphere. Applying Stokes' theorem (1.34), for $\Omega = -\frac{1}{2}$ on an octant, gives easily $\phi = -\pi/4$. This verifies the result found in §1.2.3.

Note that from Stokes' theorem (1.34), Berry curvature can be defined alternatively as the Berry phase per unit solid angle at direction $\hat{\mathbf{n}}$ on the Bloch sphere. With this, the requested for the general case, posed in §1.2.3, can be answered now : the Berry phase that accumulates a spinor when moves along a closed path on the Bloch sphere, is $-1/2$ times the solid angle subtended by the path, see Fig. 1.6b. For the special case where the path is a great (maximal) circle on the sphere, the Berry phase is $-\pi$ (since the solid angle subtended by a hemisphere is 2π). This is in consistency with the well known property of spinors that a 2π rotation changes the sign of the spinor, while to retrieve its initial state must be rotated by 4π, in contrast to the usual vectors[27].

1.3.3 Chern theorem and Chern number

Since the total solid angle subtended by a sphere is 4π, in the above example the Berry flux through the whole sphere is $\Phi = -2\pi$. But it is known from Vector Analysis that the flux of a vector through any a closed surface vanishes, $\oint_S \hat{\mathbf{n}}\, dS = 0$, unless a singularity (sink or source) is enclosed inside the surface. Therefore, since the Berry flux through the whole sphere is not zero, the spinor must poses a singularity. Above has already been mentioned that in the representation (1.43) of the spinor, there is a gauge implicitly choosen which makes the spinor smooth and continuous in the north pole, $\theta = 0$, of the sphere. But this is not the case for the whole sphere.

To reveal the singularity let start computing the Berry flux through the whole sphere, starting with a section near the north pole and gradually progressing to the south pole. For convenience, the calculation is done using a dodecahedron (12 pentagons) as a model of the sphere, Fig. 1.8.

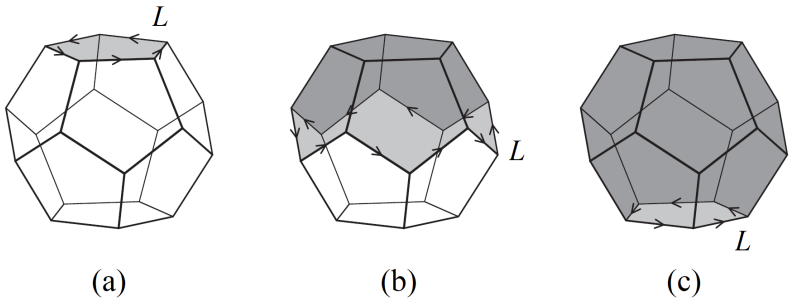


Figure 1.8: (reprinted from [252]). Calculation of the Berry phase of a spinor, on a discretized sphere. The Stokes' theorem is applied (a) on the top pentagon, (b) on the top six pentagons, (c) on all pentagons except the bottom one.

[27] Also, two spinors are peperdicular when the angle they form is π, not $\pi/2$ as the usual vectors.

Initially, the circulation of Berry connection **A** is computed along the top pentagon, Fig. 1.8a. As this is 1/12 of the whole surface, the Berry flux (hence the circulation of **A**, which is the Berry phase) is 1/12 of -2π, that is $-\pi/6$. In the same way, the circulation for a section comprised of six pentagons, Fig. 1.8b, is $-\pi$. The last case, Fig. 1.8c, is more subtle. The calculation of Berry flux through the section of the 11 pentagons results in $-11\pi/6$, as expected. But alternatively, the flux can be computed for the bottom pentagon taking its boundary to be traced in the opposite sense, or equivalently the unit normal to point outwards[28] ˙ this gives the flux is $\pi/6$, a different result. This seems to be a contradiction – but it is not. The key point is that the Berry phase is well defined only modulo 2π ˙ in this context $-11\pi/6$ and $\pi/6$ are the same value. Generalizing, for any closed surface discretized in cells, the total circulation[29] (Berry phase), equivalently the Berry flux, it is 2π times an integer.

In the continuous regime, this means that the Berry flux Φ_S computed on any closed 2D manifold equals 2π times an integer C, that is

$$\Phi_S = \oint_S \mathbf{\Omega} \cdot d\mathbf{S} = 2\pi C \,. \tag{1.45}$$

This is known as the *Chern Theorem*, and the integer C is the so called *Chern number* of the surface, also known as *Chern index* or *TKNN invariant*[30]. Chern number is a topological invariant which concerns the manifold (regarded as a surface) of the states $|u_{\boldsymbol{\lambda}}\rangle$ in the parameter space. It is stressed out that, since the integration in (1.45) involves implicitly the states $|u_{\boldsymbol{\lambda}}\rangle$, the Chern number emanates here primarily from the nature of the states, not from the geometry of the surface[31].

The Chern theorem stems mathematically from the gauge-invariance modulo 2π of the Berry phase. To give a proof of it in the continuous regime, it is necessary to clarify how this issue is manifested in Stokes' theorem, (1.34). The left-hand side (lhs) of (1.34) is the flux of Berry curvature through the surface S ˙ since $\mathbf{\Omega}$ is gauge-invariant, this term is determined uniquely. The right-hand side (rhs) is the Berry phase along the boundary L of S, which is gauge-invariant and uniquely defined only modulo 2π. Therefore, the two sides are not unconditionally equal. To remove the ambiguity, the Stokes' theorem (1.34) is interpreted in a conditionally way. Specifically, if the Berry phase is computed using information for $|u_{\boldsymbol{\lambda}}\rangle$ only on the curve L, then it is uniquely defined only modulo 2π. This is expressed restating (1.34) as

$$\Phi_S = \int_S \mathbf{\Omega} \cdot d\mathbf{S} := \oint_L \mathbf{A} \cdot d\boldsymbol{\lambda} = \phi_L \,. \tag{1.46}$$

[28] Observe that a boundary which is traced anticlockwise in the north hemisphere, as it is expanding and getting into the south hemisphere, its sence of tracing *becomes clockwise but remains positive* for the surface that is expanding from north. In contrast, its complement area in south hemisphere has the same boundary but it is traced anticlockwise (the unit normal points outwards), which is *negative for the area getting closer from north*. That is the case here.

[29] From now on "circulation" will mean the "circulation of the Berry connection **A**", unless otherwise stated.

[30] after the author names Thouless, Kohomoto, Nightingale, and den Nijs, of the important paper [242].

[31] Another topological invariant for a surface is the Euler characteristic χ, well known from Topology. χ is an integer concerning the topological structure of the surface, and is independent of the nature of any physical quantities that build the surface. For the affinity of χ with the Chern number see §1.5.

The symbol $:=$ is used here to give a special meaning in the equality : it indicates that the uniquely defined lhs quantity equals to one of the values of the rhs quantity, which is uniquely defined modulo 2π. Therefore, the meaning of (1.46) is that the equality holds if an appropriate gauge is choosen (if this is feasible) for the states $|u_{\boldsymbol{\lambda}}\rangle$ along the closed path L, while other gauges produce a difference of an integer multiple of 2π.

The equality in the expression (1.46) holds for a gauge which is smooth and continuous on the whole S, including its bounday L ˙ for the Berry phase computation in a closed path, this gauge annihilates the mismatch of integer multiple of 2π. When regarding $|u_{\boldsymbol{\lambda}}\rangle$ as a function defined only in the neighborhood of L (i.e., locally), it is feasible to make a gauge transformation that shifts ϕ_L by 2π ˙ unfortunately, such a gauge change cannot be defined for the whole interior of S (i.e., globally) without a vortex-like singularity.

With all the above, the Chern theorem (1.45) can be prooved very easily. The proof is given for a surface topologically equivalent to a aphere, and then it can be generalized for any orientable 2D surface, for example a torus. These cases (sphere & torus) are the most common in the applications.

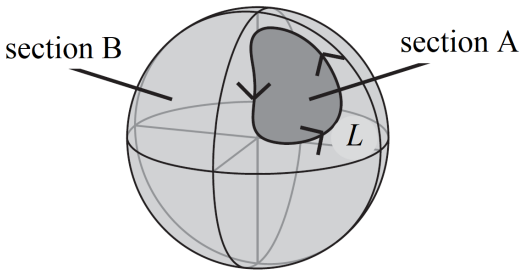


Figure 1.9: Geometry for the proof of Chern theorem.

Let a sphere with its surface divided to two sections, A and B, Fig. 1.9. The closed path L is the boundary between the two sections ˙ its traversing direction (anticlokwise) is positive for A and negative for B. Applying Stokes' theorem to A and B gives respectively $\int_A \boldsymbol{\Omega} \cdot d\mathbf{S} := \phi_L$, and $-\int_B \boldsymbol{\Omega} \cdot d\mathbf{S} := \phi_L$, where ϕ_L is the Berry phase along L. The result for ϕ_L must be the same for the two sections, but only modulo 2π. Subtraction of these two equations gives

$$\int_A \boldsymbol{\Omega} \cdot d\mathbf{S} + \int_B \boldsymbol{\Omega} \cdot d\mathbf{S} = \oint_S \boldsymbol{\Omega} \cdot d\mathbf{S} := 0 \tag{1.47}$$

which is equivalent to (1.45).

The generalization of Chern theorem to any closed, orientable 2D surface, like a torus, is quite easy, and will be discussed only briefly. The method is to setup an atlas for the surface, composed of a set of mappings[32], and in each map a smooth and continuous gauge is defined. For each map the Stokes' is applied, and the equations are summed

[32] The terms "atlas" and "map" for a surface (more correctly for a manifold) are defind strictly in the context of Differential Geometry.

by members. In one side, the sum is $\oint_S \mathbf{\Omega} \cdot d\mathbf{S}$ over the whole surface S ˙ on the other side is the sum of Berry phases along the boundaries of the maps, which are eliminated modulo 2π. Thus, the conclusion is the same as for the simple sphere examined above, and the Chern theorem has been prooved[33].

1.3.4 A simple example of Chern number computation

Continuing the example with the spinor, in §1.3.2, the $\mathbf{\Omega} = -\frac{1}{2}\hat{\mathbf{n}}$ found there gives for the whole Bloch sphere $\oint_S \mathbf{\Omega} \cdot d\mathbf{S} = -2\pi$, and this implies via (1.45) a Chern number $C = -1$. In general, for a spin-s particle it is $C = -2s$, and since the Chern theorem demands C to be an integer, it follows that the only allowed spinors are those with half-integer or integer spin. This is well known from Quantum Mechanics, verified also here. In this way, a spin-1 particle gives $C = -2$, and a spin-$\frac{3}{2}$ gives $C = -3$.

It is streesed out that when the Chern number is non-zero, construction of a gauge smooth and continuous on the whole surface S is impossible. If such a gauge could exist, then the Stokes' theorem could apply directly to the whole surface and result that the Chern number is zero, contradicting to the intial assumption. This can be demostrated in the example for the spinor discussed above. Initiallly, let $\hat{\mathbf{n}} = +\hat{\mathbf{z}}$ and define a gauge[34] smooth in a neighbourhood of $\theta = 0$. Then, extend smoothly this gauge on the sphere while increasing gradually θ. This results in a spinor representation like (1.43). But as discussed in §1.3.3, this representation has a singularity at the south pole, $\theta = \pi$. Alternativelly, the inital point could be the south pole, $\theta = \pi$, with $\hat{\mathbf{n}} = -\hat{\mathbf{z}}$. This would result in the spinor representation

$$|\uparrow_{\hat{\mathbf{n}}}\rangle = \begin{bmatrix} \cos(\theta/2)e^{-i\phi} \\ \sin(\theta/2) \end{bmatrix} \tag{1.48}$$

equivalent to (1.43). But now, although this is smooth at $\theta = \pi$, it has a singularity at the north pole, $\theta = 0$. In fact, it is impossible to define a gauge which is smooth and continuous on the whole sphere. This is manifested by the Chern number : a non-zero Chern number expresses that the topological structure of the manifold does not permit a gauge that is smooth on the whole manifold [241, 244]. Besides this, even when the Chern number is zero, to construct a gauge that is smooth at least locally (i.e., the states to be smooth functions of the parameters in only a neighborhood of the parameter space), although it is theoretically feasible [234, 267], in practice may be very difficult.

1.3.5 Berry phase and Adiabatic Dynamics

For the most physical systems, especially the quantum ones, the set of states whose Berry phase is of interest are eigenstates of some Hamiltonian $\widehat{H}(\lambda)$. In these systems,

[33] A rigorous proof of Chern theorem, in terms of Algebraic Topology and Differential Geometry, can be found in [60], [68] and [174].

[34] in fact a spinor representation, which implies a gauge implicitly chosen.

the variation of the parameter λ, and the evolution of states $|u_\lambda\rangle$, are supposed to be adiabatic[35], as already explained in §1.2.3. In essence, this means that the variation and the evolution is continuous ˙ the discussion untill know was done with this condition (i.e., adiabatic approximation) adopted silently wherever needed to ensure this continuity. In the adiabatic approximation regime, the system remains in the same eigenstate during its moving in the parameter space ˙ only the phase of the eigenstate changes. For quantum systems it is reasonable to study how this evolution and the accumulation of Berry phase is related to the evolution of the system as described by the time-dependent Schrodinger equation.

Let consider the Hamiltonian $\widehat{H}(\lambda)$ of a quantum system, with eigenstates $|n(\lambda)\rangle$, where n labels the eigenstates. It is adopted that the paramater $\lambda(t)$ is a slow[36] function of time t. It is well known that the eigenstates of $\widehat{H}$, for a given λ, satisfy the equation

$$\widehat{H}(\lambda)\,|n(\lambda)\rangle = E_n(\lambda)\,|n(\lambda)\rangle\,. \tag{1.49}$$

Initially ($t = 0$), the system is considered to be in eigenstate n and its time evolution is recorded from there on.

When λ is independent of t, the eigenstate exhibits a time dependence as [177]

$$|\psi(t)\rangle = e^{-iE_n t/\hbar}\,|n(\lambda)\rangle\,. \tag{1.50}$$

Therefore, in a small time interval Δt the phase accumulates a value $e^{-iE_n\Delta t/\hbar}$.
But let λ varies slow enough with time to can be considered constant in each interval Δt. Then, the total phase accumulated for a number of time intervals is

$$\prod e^{-iE_n\Delta t/\hbar} = e^{-i\sum E_n\Delta t/\hbar}\,. \tag{1.51}$$

In the contiuum limit, the sum in (1.51) becomes an integral, and the above is written as $e^{-i\frac{1}{\hbar}\int_0^t E_n(t')dt'}$. Setting

$$\gamma(t) = \frac{1}{\hbar}\int_0^t E_n(t')\,dt'\,, \tag{1.52}$$

the evolution of the eigenstate, (1.50), becomes $|\psi(t)\rangle = e^{-i\gamma(t)}\,|n(\lambda)\rangle$.
(1.49) concerns stationary states. However, here the eigenstates considered to change adiabatically. To find the correct variation of their phase, an additional term must be introduced in (1.50). Therefore, in the regime of adiabatic approximation, instead of (1.50), it is used the ansatz

$$|\psi(t)\rangle = c(t)\,e^{-i\gamma(t)}\,|n(t)\rangle \tag{1.53}$$

[35] In §1.2.3 the term "adiabatic evolution" was explained as "slow enough for the static approximation to be reliable". This has the meaning that the variation of λ and $|u_\lambda\rangle$ is small compared to a main characteristic (usually time-like in nature) of the system. For example, in the conductance study of a material, the variation rate of λ along a path in the parameter space, to be small compared to the frequencies corresponding to the energy gap of the material.

[36] The meaning of "slow" will be clarified below.

where the factor $c(t)$ takes account the extra phase advance (if there is any) beyond the formal caught by $\gamma(t)$.
Also, $|n(t)\rangle$ is simply the eigenstate $|n(\lambda)\rangle$ of the time-independent problem, evaluated at $\lambda = \lambda(t)$ · i.e., $|n(t)\rangle$ really is $|n(\lambda(t))\rangle$, where the $|n(\lambda)\rangle$ is stationary solution and its time variaton is adiabatic. This ansatz is in fact only the zero-order term in a perturbation expansion of $|\psi(t)\rangle$ in $d\lambda/dt$. The necessity to occasionally include power terms of higher order will be discussed later.

Setting the ansatz (1.53) in the time-dependent Schrodinger equation

$$i\hbar\partial_t|\psi(t)\rangle = \widehat{H}(t)\,|\psi(t)\rangle\,, \tag{1.54}$$

where $\partial_t = \partial/\partial t$, gives after some algebra[37]

$$\dot{c}(t)\,|n(t)\rangle + \dot{c}(t)\,\partial_t|n(t)\rangle = 0\,. \tag{1.55}$$

Multiplying (1.55) on the left with the bra $\langle n(t)|$ gives

$$\dot{c}(t) = i\,c(t)A_n(t)\,, \tag{1.56}$$

where has been set $A_n(t) = \langle n(t)|i\partial_t n(t)\rangle$.
But from (1.14) it is inferred that $A_n(t)$ is a Berry connection, with time as its parameter. Solving (1.56) gives $c(t) = e^{i\phi(t)}$ with

$$\phi(t) = \int_0^t A_n(t')\,dt' \tag{1.57}$$

which evidently is a Berry phase in an open path, with time as parameter.

Furthermore, it is desirable to express the Berry phase (1.57) in its formal form, with parameter λ instead of the time. Above, $|n(t)\rangle$ was defined to be $|n(\lambda(t))\rangle$ · thus, applying the chain rule gives $\partial_t|n(\lambda(t))\rangle = \dot{\lambda}\,\partial_\lambda|n(\lambda)\rangle$.
But then it is $A_n(t) = \dot{\lambda}A_n(\lambda)$, where $A_n(\lambda) = \langle n(\lambda)|i\partial_\lambda n(\lambda)\rangle$ is the Berry connection in parameter space. Substituting this in (1.57), and using $d\lambda = \dot{\lambda}dt$, results in

$$\phi(t) = \int_{\lambda(0)}^{\lambda(t)} A_n(\lambda)\,d\lambda\,. \tag{1.58}$$

---

[37] It is

$$\partial_t e^{-i\gamma(t)} = -\,i\frac{\partial\gamma(t)}{\partial t}e^{-i\gamma(t)}\,,$$

and differentiating (1.52) gives

$$\frac{\partial\gamma(t)}{\partial t} = \frac{1}{\hbar}E_n(t)\,.$$

The derivative ∂_t acts on all the terms of (1.53) and the term $\partial_t e^{-i\gamma(t)}$ is cancelled versus the $\widehat{H}(t)\,|n(t)\rangle = E_n(t)\,|n(t)\rangle$, ending in (1.55).
It is reminded that the states $|n(t)\rangle$ are considered to be unitary.

This result is noteworthy. It means that the Berry phase involved into the time-dependent wavefunction (1.53) is a function of only the path traversed in the parameter space, and is independent of the rate of moving on the path[38] ˙ this holds under the condition that the parametric evolution is sufficiently slow (adiabatic).

The ansatz (1.53) is correct only if the term $c(t)$ is included, which takes account the Berry phase. Therefore, in the adiabatic regime, the evolution of the wavefunction in time is[39]

$$|\psi(t)\rangle = e^{i\phi(\lambda(t))}\, e^{-i\gamma(t)}\, |n(t)\rangle \tag{1.59}$$

where the Berry phase $e^{i\phi(\lambda(t))}$ is added to the formal dynamic phase $e^{-i\gamma(t)}$.

It is emphasized again that, in contrast to the conception that "the phase does not matter", in some cases the phase can cause interference phenomena with important physical meaning and uses ˙ thus, it is wrong to naively ignore the phase always.

Also, it is pointed out that if a PT gauge is set for $|n(\lambda)\rangle$, then by definition is $A_n(\lambda) = 0$ and $\phi(t) = 0$. In this case, the Berry phase term is absent in (1.59) and the system evolves following this PT gauge.

It was mentioned earlier that (1.53) or (1.59) is only the zero order term in a perturbation expansion of $|\psi(t)\rangle$ in $d\lambda/dt$, and that in some cases it is necessary to include power terms of higher order. Such an example is the adiabatic charge transport, which is important in crystalline systems. A more supervisory, simplified case is an individual atom or molecule. It is known that the current density for an electron in state $|\psi\rangle$ is [177]

$$\mathbf{j}(\mathbf{r}, t) = \frac{ie\hbar}{2m}\big(\psi^*(\mathbf{r}, t)\nabla\psi(\mathbf{r}, t) - \psi(\mathbf{r}, t)\nabla\psi^*(\mathbf{r}, t)\big) \tag{1.60}$$

which vanishes indentically when $\psi(\mathbf{r}, t)$ is real. This holds also for the ψ given by (1.59), since its phase factors are indepented on $\mathbf{r}$. Let apply this in a typical case, e.g., for the ground electronic state of a NH_3 molecule as one nucleus moves slowly ˙ then $|n(\lambda)\rangle$ is indeed real. Adopting ψ given by (1.59) results that the motion of the nucleus does not produce any flow of electron charge. But this is evidently wrong because $\rho(\mathbf{r})$ changes with time and thus current flow must be induced.

The obviation of this contradiction is to include higher power terms of $\dot{\lambda}$ in the adiabatic approximation of ψ. In fact, one more term is enough. Thus $|\psi(t)\rangle$ in (1.59) is expanded as

$$|\psi(t)\rangle = e^{i\phi(\lambda(t))}\, e^{-i\gamma(t)}\, \big[\, |n(t)\rangle + \dot{\lambda}|\delta n(t)\rangle \,\big] \tag{1.61}$$

where the additional term $|\delta n(t)\rangle$ must be determined.
Above has been found that (1.61) is the solution to the time-dependent Schrodinger equation (1.54) at zero order in $\dot{\lambda}$. Here it is requested to be a solution at first order

[38] This important property also explains the alternative name "geometric phase" for the Berry phase, as it reminds that this phase depends only on the path in the parameter space and not on the velocity of moving on it or the representation of the states on the path.

[39] It is again pointed out that this is only the main term (zeroth order) of a perturbation expansion of ψ in $d\lambda/dt$.

too. To this end, (1.61) is put in (1.54), terms higher than first order ($\ddot{\lambda}$, $\dot{\lambda}^2$, $\dot{\lambda}\partial_t|\delta n(t)\rangle$ etc) are discarded and the remain is

$$(E_n - \widehat{H}_\lambda)\,|\delta n\rangle = -i\hbar(\partial_\lambda + iA_n)\,|n\rangle\,. \tag{1.62}$$

Since it is $A_n(t) = \langle n(t)|i\partial_t n(t)\rangle$, and setting[40] $Q_n = 1 - |n\rangle\langle n|$,
the above is written as

$$(E_n - \widehat{H}_\lambda)\,|\delta n\rangle = -i\hbar Q_n\,|\partial_\lambda n\rangle\,, \tag{1.63}$$

from which the $|\delta n\rangle$ can be found.
Specifically, the solution to (1.63) is [252]

$$|\delta n\rangle \;= -i\hbar T_n^2(\partial_\lambda \widehat{H})|n\rangle \tag{1.64}$$

where $T_n = \sum_{m\neq n} \frac{|m\rangle\langle m|}{E_m - E_n}$.
Eventually, (1.64) can be expressed as a sum over the eigenstates $|m(\lambda)\rangle$, as [252]

$$|\delta n\rangle \;= -i\hbar \sum_{m\neq n} \frac{\langle m|(\partial_\lambda \widehat{H})|n\rangle}{(E_m - E_n)^2}|m\rangle\,, \tag{1.65}$$

which was the requested.

It is pointed out that $|\delta n\rangle$ was found to be independent of time. This means that the evolution of the wavefunction as seen in (1.61) is mainly due to λ, assisted by the term with $\dot{\lambda}$ [41].

It can be prooved that if (1.61) is used to compute the expectation value of the current operator (1.60), the correct description for the the charge transport during the adiabatic evolution is obtained.

All the above study holds in the adiabatic regime, which ensures that the variation of λ and the eigenstates are continuous. It is purposeful to quantify this condition, even roughly. The concept is to define the "slowness" by compare the variation rate $\dot{\lambda}$ to the frequencies corresponding to the energy gap ˙ this variation must be small enough so that the adiabatic approximation holds. To this end, consider the energy levels of a quantum system and let ΔE be a typical gap between them. For two neighbouring eigenstates $|n\rangle$, $|m\rangle$ of the system, define $\lambda_0 \equiv 1/\langle m|\partial_\lambda|n\rangle$. λ_0 can be interpreted as the scale of λ over which $|n\rangle$ varies significantly. Then, for an estimation to the order of magnitude, the quantity $\hbar\dot{\lambda}/\lambda_0\Delta E$ can be indetified as a dimensionless parameter which quantifies the "slowness" of the adiabatic evolution. Practically, the evolution is adiabatic if $\dot{\lambda}$ is small compared to $\Delta E/\hbar$, i.e., to the frequency characterizing the quantum behaviour of the system.

Another interesting feature of the adiabatic regime is that, except the phase information recorded in the Berry phase, the time dependent wavefunction has only a "short

[40] i.e. Q_n is the complement of the projector operator $|n\rangle\langle n|$.

[41] See [243] for more on this.

memory" of the evolution on the path. This means that the evolution in the present time depends only from a small and very recent part of the past of the history on the path. For example, keeping terms to first order in $\dot{\lambda}$ in the state representation, the state at time t depends only on H_λ at time t and on an infinitesimally time interval prior to t. The more higher order terms are included, the more the "memory" increases (i.e., the dependence from greater past time intervals) ˙ however, the general behaviour of the system is that its evolution does not affected from what happens in earlier times during its moving along the path.

The adiabatic regime can be successfully adopted for a system which has variables with very different different variation rates (time scales). A typical example is a system of an atom or crystalline solid, with its electrons. An electron is many orders of magnitude lighter than the nuclei. For the evolution in time of such a system, the coordinates $\mathbf{r}_j$ of the nuclei are considered to be classical variables that evolve slowly on a path. But there is also a back-reaction on the system of nuclei, such that in a quantum treatment they undergo a "gauge potential" $\mathbf{A}_j$ caused the electron(s) system as it follows adiabatically the nuclear one. This potential is in fact the Berry connection $A_{jm} = \langle \psi_{\mathbf{r}} | i \frac{\partial \psi_{\mathbf{r}}}{r_{jm}} \rangle$, where $\psi_{\mathbf{r}}$ is the ground state of electrons at a fixed $\mathbf{r}$. The application of Berry phase theory to such systems has been prooved very usefull in Molecular Physics[42].

1.4 Chern number on electronic energy bands

1.4.1 The foundation of Band-structure Theory

The electronic states in crystalline materials, neglecting their interactions, can be found from a single-particle (in this case an electron) Hamiltonian $\widehat{H}(\mathbf{k})$, which is a smooth function of the wavevector $\mathbf{k}$ of the crystal. Specifically, the solutions of the time-independent Schrodinger equation

$$\widehat{H}(\mathbf{k}) \, |\psi_{n,\mathbf{k}}\rangle = E_{n,\mathbf{k}} \, |\psi_{n,\mathbf{k}}\rangle \tag{1.66}$$

are the feasible electronic states.
Having adopted a periodic potential for the crystall, the resulting eigenstates $|\psi_{n,\mathbf{k}}\rangle$ are modulated plane waves, i.e.,

$$|\psi_{n,\mathbf{k}}(\mathbf{r})\rangle = e^{i\mathbf{k}\cdot\mathbf{r}} \, |u_{n,\mathbf{k}}(\mathbf{r})\rangle \tag{1.67}$$

where the modulation function $|u_{n,\mathbf{k}}(\mathbf{r})\rangle$ has the periodicity $\mathbf{r}_l$ of the lattice[43],
$|u_{n,\mathbf{k}}(\mathbf{r} + \mathbf{r}_l)\rangle = |u_{n,\mathbf{k}}(\mathbf{r})\rangle$. This result is known as *Bloch's theorem*,
and the wavefunctions $|\psi_{n,\mathbf{k}}\rangle$ and $|u_{n,\mathbf{k}}\rangle$ are called the *Bloch waves* or *Bloch states* of an electron, where the $|u_{n,\mathbf{k}}\rangle$ are the cell-periodic ones. The eigenvalues $E_{n,\mathbf{k}}$ are the so called *energy bands*, where n is an index indicating their sequence. It can be proved

[42] E.g., see [28].

[43] i.e., $\mathbf{r}_l$ consists of multiples (l_1, l_2, l_3) of the three basis vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ of the real-space lattice.

[103] that the Bloch waves whose wavevectors differ by a reciprocal lattice vector are identical, and the energy eigenvalues $E_{n,\mathbf{k}}$ are a periodic function of the wavevectors[44] $\mathbf{k}$ of the Bloch waves˙ these properties stem from the strict periodicity of the lattice potential and the form (1.67) of the Bloch states.

In the above way, the one-electron states of a periodic potential can be represented by energy surfaces $E = E_n(\mathbf{k})$, each one being a periodic function of the wavevector in the $\mathbf{k}$-space. These energy surfaces all together constitute the *electronic band-structure* of the crystal. Since both $|\psi_{n,\mathbf{k}}(\mathbf{r})\rangle$ and $E_n(\mathbf{k})$ are periodic in $\mathbf{k}$-space, it is sufficient to know these functions for $\mathbf{k}$ values in only the first Brillouin zone˙ their values in the whole $\mathbf{k}$-space can easily be found by a simple periodic expansion.

The bulk properties of the materials are governed mainly by their band-structure. For example, when the bands occupied by electrons and the empty ones are separated by an energy gap, Fig. 1.10a, the material is an insulator˙ if, instead, there are overlapped bands, Fig. 1.10b, the material is a conductor.

However, categorizing the materials simply by their band-structure, does not catch all their physical properties. In specific, regarding the geometric scheme[45] of the Bloch states $|\psi_{n,\mathbf{k}}\rangle$, and mainly its topological characteristics, a whole bunch of material types arises, in the context of topological classification.

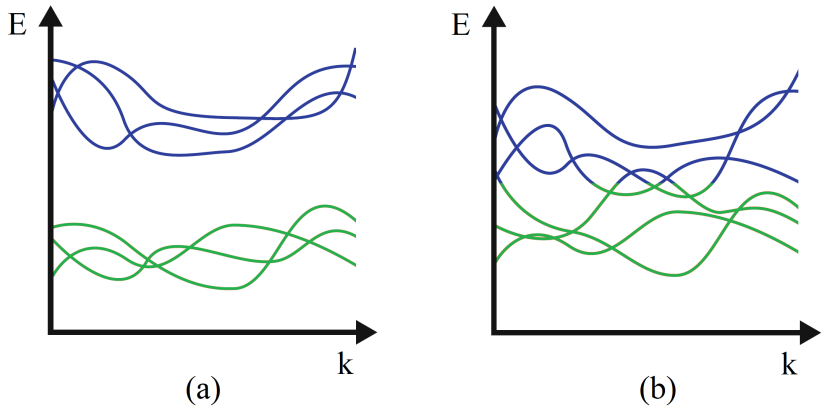


Figure 1.10: Band-structure of (a) an insulator, (b) a conductor.

1.4.2 The concept of Topological properties of materials

Many physical parameter spaces have typical geometric schemes, e.g., a cylinder or a torus. This scheme has physical significance, i.e., expresses physical properties. Typical example is the Brillouin zone of a 2D lattice representing a solid crystalline material, where the momentum vectors (k_x, k_y), $(k_x + 2\pi, k_y)$, and $(k_x, k_y + 2\pi)$ are equivalent˙ its topology is that of a torus and expresses the periodicity of the lattice.

[44] In fact, $\mathbf{k}$ (or alternative k_x, k_y, k_z) and n are quantum numbers for the eigenstates and the eigenenergies.

[45] torus, cylinder, infinite plane etc.

It is useful to remind briefly the concept of the topological properties in mathematics, so to grasp easier the corresponding concept for the properties of the materials. Very roughly said, Topology is a kind of geometry without the concept of distance[46]. Topology concerns the properties of objects – of spaces in general – which are invariant under invertible deformations (homeomorphisms[47] technically speaking). In contrast to Geometry, topologically equivalent objects resemble each other in a qualitative sense : two objects are considered equivalent if they can be deformed continuously one to another through bending, twisting, stretching, and shrinking, while avoiding tearing apart or gluing parts together. The geometric form and size of the objects is indifferent in Topology (since distance, angle and curvature do not matter) ˙ instead, what matters is, for example, if the object is connected, or if it has holes. An important, well known topological property of an object, intuitively related to the number of its holes, is the so called *genus*[48] [124]. The genus is conserved under smooth deformations of a surface, and can be used to classify surfaces by their number of holes. The only way to create a new hole or eliminate an existing one, is by tearing or gluing the surface. The genus is a topological invariant – a quantized integer that cannot be changed without changing the topological structure[49]. In general, ascribing geometric objects to physical properties or phenomena of the materials[50], topological invariants can be extracted for these objects and then used to identify and classify the materials as to these phenomena – hence the name "topological phases".

Considering the properties of the materials emanating from their band-structure, to define topological phases, it is needed a geometric object on which the topological properties can be defined. For this purpose, a set of bands B is selected. The most usual choice for B is to select the occupied subspace[51]. For each $\mathbf{k}$, the set of states $\{|u_{n,\mathbf{k}}\rangle\}_{n\in B}$ span a vector space $V_{\mathbf{k}}$ over $\mathbb{C}$. Assuming that $V_{\mathbf{k}}$ is a smooth function of $\mathbf{k}$ and the space where $\mathbf{k}$ itself is defined is a manifold[52], this defines a so called *fiber bundle*.

A simple example of a fiber bundle is given by an 1D vector space defined on a circle. If the vector space is orthogonal to the plane where the circle lies, the resulting object is a cylinder, Fig. 1.11a. Alternatively, if the basis vector is rotated gradually as it

[46] This is not fully correct since the metric spaces (which do have a metric, i.e., "distance") are a subset of the topological spaces (which in general they have not), but is acceptable for a rough description of what is Topology as a mathematical branch.

[47] Continuous invertible transformations, which have continuous inverse too.

[48] More precisely, the genus of a surface is an invariant which counts the number of tori or handles consisting the surface (for the orientable case), or the number of twisted pairs or projective planes (for the non-orientable case) [124]. It is also closely related to the Euler characteristic, mentioned earlier.

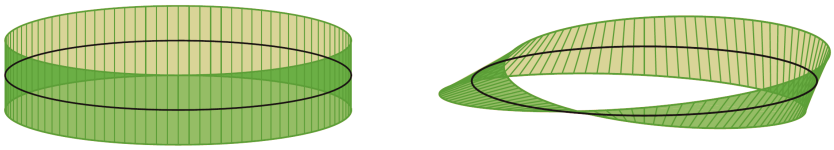
[49] Very roughly, the "topological structure" of an object is the way its fundamental parts are connected.

[50] like the torus expressing the periodicity of a 2D lattice of a solid crystalline, as mentioned above.

[51] This is not always possible ˙ an example is the semimetals, where the occupation number varies with $\mathbf{k}$. In these cases, the N lowest energy bands are selected.

[52] A *topological manifold* is a topological space of Hausdorff type, second countable, and locally homeomorphic to $\mathbb{R}^n$. The topology of a manifold is in general different from that of a vector space, and it cannot be covered by a single coordinate system. Intuitively, a manifold is constructed by pasting together many pieces of $\mathbb{R}^n$. A *differentiable manifold* is a topological manifold equipped with a differentiable *atlas*. Differentiable manifolds are a generalisation of surfaces. However, unlike the surfaces, it is not needed to consider a manifold as being immersed in a higher-dimensional space in order to study its geometric properties. More on manifolds and their associated stuff can be found in books on Differential Geometry, e.g. [68].

marches around the circle, and has been rotated by π when completes exactly one round, the resulting object is a Mobius strip, Fig. 1.11b. These two objects are topologically different because they cannot be transformed into each other via an homeomorphism.



(a) (b)

Figure 1.11: Fiber bundles. (a) A cylinder, spanned by a vector which does not rotate as it marches around a circle. (b) A Mobius strip, spanned by a vector which rotates by π as it marches around a circle.

It is clarified that the topological classification of the materials does not concern the geometrical shape of the crystal in real space, nor the shape of the Brillouin zone. Instead, *it concerns the way the states change as a function of* $\mathbf{k}$ *in the Brillouin zone*. As it is emphasized below, for the topological consideration of materials, the theory is built using the periodic Bloch functions $|u_{n,\mathbf{k}}\rangle$ and a $\mathbf{k}$-dependent Hamiltonian $\widehat{H}(\mathbf{k})$. In terms of the kinetic energy operator and the Coulomb potential, the form of the Hamiltonian is the same for all materials. What differentiates it from case to case it is the division of the eigenstates to occupied and unoccupied states[53]. The topological consideration concerns the topology of the occupied states[54] of $\widehat{H}(\mathbf{k})$ as a function[55] of $\mathbf{k}$. The variation of the eigenvalues and eigenfunctions depends on the details, but no matter how complicated the problem is, a system can be classified by a topological invariant.

In such a system, the only way the topology can change is an energy bandgap to vanish (close). The reason is the following. In the Brillouin zone, at the points where the gap between occupied and unccupied states closes, there are eigenstates of $\widehat{H}(\mathbf{k})$ that touch or even overlap, becoming degenerate. Then, the occupied and unoccupied states can exchange eigenfunctions, changing in this manner the way the set of filled eigenstates are connected in the Brillouin zone. In an insulator there is a gap between the filled and the empty bands, and all systems that can be transformed one into another by varying continuously the Hamiltonian, without a gap closing, are considered to have the same topology. Transitions between insulating states with different topologies can take place only if a gap closes.

---

[53] This can be formulated using the matrix elements of $\widehat{H}(\mathbf{k})$.

[54] More explanatory, their grouping and behaviour in the topological sense.

[55] The $\mathbf{k}$-dependent Hamiltonian for the Bloch periodic functions $|u_{n,\mathbf{k}}\rangle$ can be formulated by the usual Hamiltonian by the relation [252]

$$\widehat{H}(\mathbf{k}) = e^{-i\,\mathbf{k}\cdot\mathbf{r}}\,\widehat{H}\,e^{i\,\mathbf{k}\cdot\mathbf{r}},$$

such that $\widehat{H}(\mathbf{k})\,|u_{n,\mathbf{k}}\rangle = E_{n,\mathbf{k}}\,|u_{n,\mathbf{k}}\rangle$.

1.4.3 The Bulk-Edge Correspondence

Above, it was pointed out that the vector space $V_{\mathbf{k}}$ must be a smooth function of $\mathbf{k}$. As will be explained right now, this has a great influence on the physical properties of topological phases.

Even if the Hamiltonian $\widehat{H}(\mathbf{k})$ is a smooth function of $\mathbf{k}$, this does not necessarily hold for $V_{\mathbf{k}}$. The following example is an indicative case. Let the Hamiltonian

$$\widehat{H}(k) = -\cos(k)\,|a\rangle\langle a| + \cos(k)\,|b\rangle\langle b| \tag{1.68}$$

where $|a\rangle$, $|b\rangle$ are two arbitrary orhogonal eigenstates.
For $k = 0$, the eigenvalues of $|a\rangle$ and $|b\rangle$ are repsectively -1 and 1. Let name the energy band of $|a\rangle$ by 1 and the energy band of $|b\rangle$ by 2, where these are simply indices, 1 indicating the lowest eigenvalue and 2 the next one. As k varies, the eigenvalues also vary, until $k = \pi/2$, where they are equal. At this value, the vector space $V_{\mathbf{k}}$ switches from being spanned by $|a\rangle$, $V_{\mathbf{k}} = \text{span}(\{|u_{n,\mathbf{k}}\rangle\}_{n\in\{1\}})$, to being spanned by $|b\rangle$, $V_{\mathbf{k}} = \text{span}(\{|u_{n,\mathbf{k}}\rangle\}_{n\in\{2\}})$. Since this switching produces a discontinuity for $V_{\mathbf{k}}$, this space does not satisfy the condition for topological classification.

The smoothness of the vector space $V_{\mathbf{k}}$ can break if the order of eigenvalues between the states which belong to the set B and those which do not, change. This can be avoided easily if the possible selection of bands B is ristricted in a way that they are always separated from the other bands by a sufficient energy gap. This means that the topological properties must be searched and extracted for isolated sets of bands, which form smooth fiber bundles.

An equivalent way to set this requirement (i.e., defining topological properties only in isolated energy bands), is by looking for the transformations which can be done to a material without changing its topological properties. To the requirement that these transformations must change the Hamiltonian definitely smoothly, it is imposed additionally that the band gap must remain open. This definition for the admissible transformations of a material leads to a noteworthy physical property of topological phases : at the boundaries of topologically non-trivial insulating materials, stable conducting edge states do form. In specific, at the interface where the Hamiltonian gradually interpolates between two insulating states with different topologies, the following happens. At some point the energy gap has to close because otherwise it is impossible for the topological invariant to change. Therefore, as the gap tends to zero and finally closes, low-energy electronic states bound to the interface region appear, and these states form bands that propagate along the interface. This phenomenon is known as the *bulk-edge correspondence*, and its variations produce the interesting transport phenomena which take place in many topological materials [90, 116, 117].

The relation of topology and gapless states appears in many cases in physics, such as the gradual interpolation between regions described by Dirac Hamiltonians with positive and negative masses, which can be solved analytically [104, 95]. The most famous case of bulk-edge correspondence is the edge state in the quantum Hall effect˙ other cases are solitons in one dimension at a boundary in the Su-Schrieffer-Heeger (SSH) model for polyacetalene [238], and Majorana modes at the surface of superconductors, all of which are closely related to the so called Shockley transition in the bulk and surface states.

Despite the significance of the surface states, it is emphasized that the topology is a property of the states as a function of **k** in an infinite periodic crystal with no surfaces. It is the topology arguments that are the most basic because they ensure the robustness of the results, i.e., the qualitative conclusions will not change due to changes in the Hamiltonian, as far as the topology does not change and the system remains an insulator with a gap in its electronic structure.

1.4.4 Berry phase and Chern number of the electronic bands

It has already been mentioned the relation (1.41) for the Berry curvature of a quantum system, given by Berry himself [21]. Besides this, all the theory for the Berry phase and Chern number, studied in a general context in previous Sections, can be particularized for the electronic bands of materials. This is done briefly here and in the next Sections.

For reasons that will not be mentioned here, it is strongly emphasized that, concerning the electronic bands, the Berry phase and related quantities must be defined in terms of the cell-periodic functions $|u_{n,\mathbf{k}}\rangle$, not in terms of the Bloch functions $|\psi_{n,\mathbf{k}}\rangle$. Very briefly speaking, when using $|\psi_{n,\mathbf{k}}\rangle$, the computation of the inner product integrals in the Brillouin zone is problematic (gives zero or depends on the cell location). Instead, this does not happen with $|u_{n,\mathbf{k}}\rangle$. Furthrmore, using $|u_{n,\mathbf{k}}\rangle$ brings a deeper, more important consequence. All the $|u_{n,\mathbf{k}}\rangle$ have the same periodic boundary condition; e.g., in an 1D case imposing $u_{n,\mathbf{k}}(x=0) = u_{n,\mathbf{k}}(x=a)$, where a is the lattice constant. Therefore, all the $|u_{n,\mathbf{k}}\rangle$ belong to the same Hilbert space. As a result, inner products between vectors at different **k**, or derivatives with respect to **k**, are well defined. This would not hold if the formalism were based on the Bloch functions $|\psi_{n,\mathbf{k}}\rangle$.

Also, it is adopted the assumption that in the whole Brillouin zone the electronic bands are isolated, i.e., a band n is not overlapped with its neighbouring $n \pm 1$ bands. This restriction is important because such overlappings are common at points of high symmetry in the Brillouin zone of crystalline materials. These points of degeneracy introduce a non-analytic depedence of $|\psi_{n,\mathbf{k}}\rangle$ on **k**, making problematic the definitions of Berry connection and curvature. This problem can be amended [252] but the relevant theory is quite complicated and it will not be considered here.

Adopting the wavenumber **k** of the lattice of a material as the parameter space, the Berry phase (1.31) for the electronic bands is defined to be[56] [277]

$$\gamma_{_L} = i \oint_L \sum_{n\in B} \langle u_{n,\mathbf{k}} | \nabla_{\mathbf{k}} | u_{n,\mathbf{k}} \rangle \cdot d\mathbf{k} \tag{1.69}$$

where L is a closed path in the reciprocal space.

[56] For the Berry phase on the bands, to differentiate it from the general, the symbol γ is used instead of ϕ. For conciseness, the total Berry phase (i.e., for all bands) is considered here. The Berry phase can also be considered for a single band; in that case the sum over bands is dropped.

In this case, the Berry connection and curvature, (1.29) and (1.32), are respectively

$$\mathbf{A}(\mathbf{k}) = i \sum_{n \in B} \langle u_{n,\mathbf{k}} | \nabla_{\mathbf{k}} | u_{n,\mathbf{k}} \rangle \tag{1.70}$$

$$\mathbf{\Omega}(\mathbf{k}) = \nabla_{\mathbf{k}} \times \mathbf{A}(\mathbf{k}) \tag{1.71}$$

and the Stoke's theorem (1.46) is rewritten as

$$F_{_S} = \int_S \mathbf{\Omega} \cdot d\mathbf{S} := \oint_L \mathbf{A} \cdot d\mathbf{k} = \gamma_{_L} \tag{1.72}$$

where $F_{_S}$ denotes here the Berry flux through a surface S on the reciprocal space of the latice. The Chern number is given formally by (1.45) as [242]

$$C = \frac{1}{2\pi} \oint_S \mathbf{\Omega}(\mathbf{k}) \cdot d\mathbf{S}\,. \tag{1.73}$$

As was discussed thoroughly in §1.3.3, in contrast to common intuition, the Chern number for a closed surface can be non-zero. For the equality in (1.72) to hold, the Berry connection $\mathbf{A}(\mathbf{k})$ must be smooth. But this is not sufficient to guarantee a zero Chern number; the inherent nature of states in the parameter space can prohibit a globally smooth gauge, making the Berry connection non-smooth in the whole surface and giving it a winding value (i.e., a non-zero Chern number) in a closed path.

As far as the electronic bands concerns, in §1.4.3 it was emphasized that for $V_{\mathbf{k}}$ spanned by $|u_{n,\mathbf{k}}\rangle$ to be susceptible to topological classification, $V_{\mathbf{k}}$ must be a smooth function of $\mathbf{k}$. However, this does not impose the Berry connection for the states to be definitely smooth; as a result, the Chern number in the reciprocal space of a lattice can in general take integer, non-zero values.

Furthermore, the Chern number of a band of an insulator is a topological invariant in the following context. The Hamiltonian describing the system of electrons of the lattice can be considered to be deformed adiabatically (thus continuously) and maintaining open the energy gaps that separate the nth band from the other bands. In such a situation, the Berry curvature varies continuously; consequently, its integral over the Brillouin zone (i.e., the Berry phase), which equals 2π times the Chern number, cannot change since the Chern number is necessarily an integer; therefore, the Chern number cannot change. In contrast, if the Hamiltonian is deformed non-adiabatically (hence non-continuously), then some energy gap separating the nth band from a neighboring band might close and reopen, and the Chern number might change. In this context, the Chern number for 2D lattice models is a topological invariant like the winding number is for the 1D SSH model [7].

Lastly, before examine more practically how to compute the Berry phase and Chern number in the Brillouin zone, an important subtlety must be mentioned. For the Berry phase to be well defined, $|u_{n,\mathbf{k}}\rangle$ must be a smooth function of $\mathbf{k}$ in the whole path, open or closed. For example, consider a loop in an 1D band structure, parametrized by k, with $0 \leq k \leq 2\pi/a$. The Bloch function $|\psi_{n,\mathbf{k}}\rangle$ must be smooth across the artificial boundary point where k returns to 0 from $2\pi/a$.

This means that $\psi_{n,k}(x)$ must satisfy the boundary condition[57]

$$\psi_{n,k=2\pi/a}(x) = \psi_{n,k=0}(x)\,. \tag{1.74}$$

But then, the Bloch functions at the two ends of the interval $[0,\, 2\pi/a]$ must be exactly equal, with definitely the same phase. As the Berry connection must be defined using the cell-periodic functions $|u_{n,\mathbf{k}}\rangle$, remembering (1.67), the condition (1.74) is written as

$$u_{n,k=2\pi/a}(x) = e^{-i\,2\pi x/a}\, u_{n,k=0}(x)\,. \tag{1.75}$$

As a result, the functions $|u_{n,k=2\pi/a}\rangle$ and $|u_{n,k=0}\rangle$ are not equal˙ they differ in phase more than the expected Berry phase – and in fact by the factor $e^{-i\,2\pi x/a}$ which depends on x. When calculating the Berry phase using the cell-periodic functions $|u_{n,\mathbf{k}}\rangle$, this factor must definitely be taken account. For example, let calculate the Berry phase for a loop in an 1D Brillouin zone. The loop is discretized in N equal intervals, with $k_j = 2\pi j/N$, $j = 0,\, 1,\, ...,\, N-1$, and (1.1) is to be used. But to take account the extra phase for the functions $|u_{n,k_j}\rangle$, (1.1) must be used in the slightly modified form

$$\phi = -\mathrm{Im}\ln\left[\,\langle u_{n,k_0}|u_{n,k_1}\rangle\langle u_{n,k_1}|u_{n,k_2}\rangle ... \langle u_{n,k_{N-1}}|e^{-i\,2\pi x/a}|u_{n,k_0}\rangle\,\right]. \tag{1.76}$$

In (1.76) the factor $e^{-i\,2\pi x/a}$ is introduced[58] in the last inner product to get the correct phase difference between[59] $|u_{n,k_N}\rangle$ and $|u_{n,k_0}\rangle$.

1.4.5 Fundamentals of computing Chern number arithmetically from Berry phase

In §1.4.4, the Berry phase and the Chern number were expressed in terms of the cell-periodic states $|u_{n,\mathbf{k}}\rangle$. With small modifications, these formulas can be used to calculate numerically the Chern number on the Brillouin zone˙ the fundamentals will be presented briefly here[60]. For convenience, a very simple example will be used : let be the surface S for the calculation is the Brillouin zone $\mathbf{k} \in [0,\,1) \times [0,\,1)$ of a 2D lattice, where $\mathbf{k}$ is expressed in reduced coordinates, Fig. 1.12a. The technique can be applied to other closed 2D surfaces in the same way.

---

[57] In fact, (1.74) is the so called *periodic gauge condition*,

$$|\psi_{n,\mathbf{k}+\mathbf{G}}\rangle = |\psi_{n,\mathbf{k}}\rangle,$$

where $\mathbf{G}$ is a reciprocal lattice vector, that relates two states at the boundary of cells. The above is special case of the more general condition

$$|\psi_{n,\mathbf{k}+\mathbf{G}}\rangle = e^{-i\beta(\mathbf{k})}\,|\psi_{n,\mathbf{k}}\rangle,$$

which holds for states at the boundary of cells in the Brillouin zone.

[58] In general, the factor is $e^{-i\,\mathbf{G}\cdot\mathbf{r}}$, where $\mathbf{G}$ is the reciprocal lattice vector of the periodic gauge.

[59] The functions $|u_{n,k_i}\rangle$ are taken from a process (usually a matrix diagonalization routine) without caring their phase, thus it is not expected that $|u_{n,k_N}\rangle$ and $|u_{n,k_0}\rangle$ will satisfy (1.75).

[60] These formulas are used in the code package Z2Pack [84].

Firstly, the surface S is divided in sufficient small patches S_i. The integral (1.73) for the Chern number on the whole surface is given as the sum of integrals on all the patches,

$$C = \sum_i C_{S_i} , \tag{1.77}$$

where C_{S_i} are the integrals on the patches,

$$C_{S_i} = \frac{1}{2\pi} \int_{S_i} \mathbf{\Omega}(\mathbf{k}) \cdot d\mathbf{S} . \tag{1.78}$$

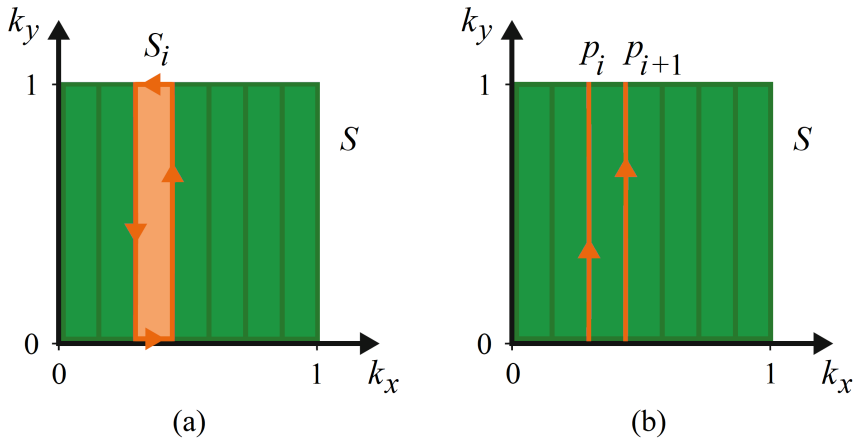


Figure 1.12: Computing the Chern number in Brillouin zone, reprensented by a surface S in $\mathbf{k}$ space (in reduced coordinates). (a) The surface is divided to patches S_i. (b) Only the path segments p_i, which traverse the Brillouin zone at constant k_x, contribute to the calculation ˙ the top and bottom ones on boundary of each patch are cancel out.

The parches S_i are small enough for the $\mathbf{A}(\mathbf{k})$ to can be made locally smooth [234, 267]. Threfore, the Stoke's theorem can be used and gives

$$C_{S_i} \mod 1 = \frac{1}{2\pi} \int_{\partial S_i} \mathbf{A}(\mathbf{k}) \cdot d\mathbf{k} \mod 1 = \frac{\gamma_{\partial S_i}}{2\pi} \mod 1 , \tag{1.79}$$

where the modulus stems from the fact that the Berry phase is uniquely defined only modulo 2π. Each C_{S_i} is much smaller than unity[61], thus, its value can be determined uniquely from $\gamma_{\partial S_i}/2\pi$ adding an integer that minimizes its absolute value.

[61] since $\mathbf{k}$ sweeps the unit square and the patches have been imposed to be sufficiently small.

Furthermore, the top and bottom parts of ∂S_i are cancel out due to periodicity[62] and the Berry phase eventually is

$$\gamma_{\partial S_i} = \gamma_{p_{i+1}} - \gamma_{p_i} \tag{1.80}$$

where p_{i+1} and p_i are the paths at either side of patch S_i, Fig. 1.12b.

Moreover, since each path p_i is on fixed k_x, the Berry phase can be considered to be a function of k_x. As both γ and k_x are periodic, the Berry phase traces a line on a torus[63], Fig 1.14. The Chern number is just the winding number of this line around the torus [233]. This means that the Chern number can be calculated tracking continuously the Berry phase on lines of constant k_x that traverse the Brillouin zone. In practice, enforcing this continuity is difficult and special care must be taken to achieve this when programming the method[64].

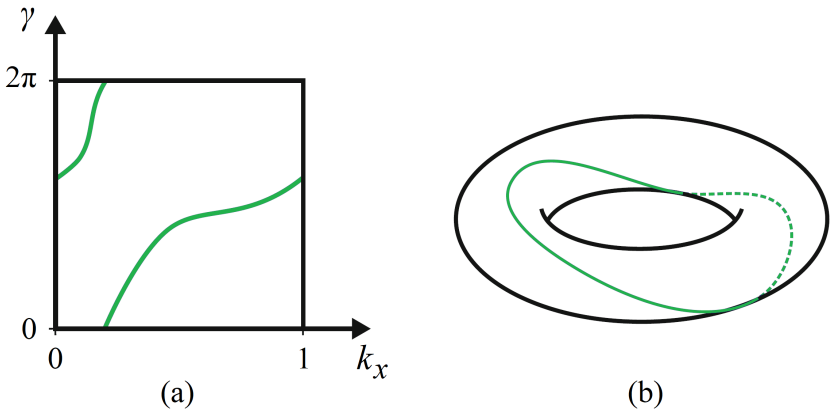


Figure 1.13: (a) Berry phase γ as a function of k_x, in Brillouin zone. The Chern number here is $C = 1$. (b) The Chern number as a winding number (here 1) on the torus geometry of the 2D lattice.

From the above, it is evident that to calculate the Chern number it must in fact to calculate the Berry phase for closed paths in the Brillouin zone. The methods used for this are the *Wilson loop* and the *Hybrid Wannier Charge Centers*, and will be presented briefly here.

---

[62] Such a cancelation does not necessarily happens on the path segments p_i that cross the Brillouin zone.

[63] See also the remark in the start of §1.4.2.

[64] For example, the code package Z2Pack [84] has a whole bunch of parameters to control the convergence due to this continuity task.

1.4.5.1 Calculating the Berry phase in Brillouin zone with the method of Wilson Loop

The *Wilson loop* $\mathbf{W}(L)$ [2] is an operator, represented by a matrix, that maps the states at a starting point $\mathbf{k}_0$ along a loop L onto their images after parallel transporting them along L. The path L is discretized into a set of points $\{\mathbf{k}_0, ..., \mathbf{k}_{n-1}, \mathbf{k}_n = \mathbf{k}_0\}$ and the Wilson loop is approximated as [84, 2]

$$\mathbf{W}(L) = \mathbf{M}_{\mathbf{k}_0,\,\mathbf{k}_1} \cdot ... \cdot \mathbf{M}_{\mathbf{k}_{n-1},\,\mathbf{k}_n} \tag{1.81}$$

where

$$M_{\mathbf{k}_i,\,\mathbf{k}_j\,mn} = \langle u_{m,\mathbf{k}_i} | u_{n,\mathbf{k}_j} \rangle \tag{1.82}$$

are overlap matrices between Bloch states at different $\mathbf{k}$.
Each eigenvalue λ_i of the Wilson Loop has as argument the rotation angle that is acquired by an eigenstate of the $\mathbf{W}(L)$ as it marches along the path L. Consequently, the total Berry phase is given via the eigenvalues of the Wilson loop as [146]

$$\gamma_{_L} = \sum_i \arg \lambda_i \,. \tag{1.83}$$

Since the overlap matrices $\mathbf{M}_{\mathbf{k}_i,\,\mathbf{k}_j}$ can be computed quite easily, the above consitute a method for calculating the Chern number numerically. However, the convergence of the Wilson loop eigenvalues relative to the discretization of L is sensitive and must be taken into account in programming the method[65].

1.4.5.2 Calculating the Berry phase in Brillouin zone with the method of Hybrid Wannier Charge Centers

The *Hybrid Wannier Charge Centers* [231, 232] provide another method to calculate the Berry phase. The foundation of this method is the *Wannier orbitals*, which are defined as the Fourier transform of the Bloch states :

$$|\mathbf{R}_n\rangle = \frac{V}{(2\pi)^d} \int_{BZ} e^{-i\mathbf{k}\cdot\mathbf{R}} |\psi_{n,\mathbf{k}}\rangle \, d\mathbf{k} \tag{1.84}$$

where $d = 1$, 2 or 3 stands for the space dimensionality, V is the unit cell volume, and the integral is taken over the first Brillouin zone. In contrary to the spreaded nature of the Bloch waves, the Wannier orbitals are localized. Also, if the Bloch waves change by a gauge transform applied on them, the Wannier orbitals change too. The choice of gauge

[65] This is just an indirect appearance of the difficulty in tracking continuously the Berry phase on lines of constant k_x, which mentioned at the end of §1.4.5.

has great influence to the properties of Wannier orbitals, especially to their localization and position in real space [169].

Furthermore, for purposes of calculating topological invariants, the *hybrid Wannier orbitals* are defined [232, 215] ˙ these are simply the Fourier transform of the Bloch states in only one spatial direction, while in the other directions are untouched and remain spreaded [84] :

$$|l_x, k_y, k_z\,; n\rangle = \frac{a_x}{2\pi}\int_{-\pi/a_x}^{\pi/a_x} e^{-ik_x l_x}|\psi_{n,\mathbf{k}}\rangle\, dk_x \tag{1.85}$$

where $l_x \in \mathbb{Z}$ and a_x is the lattice constant along the x-direction.
An hybrid Wannier orbital is localized in only one direction ˙ its average position is a function of the non-transformed variables in the reciprocal space :

$$\bar{x}_n(k_y, k_z) = \langle 0, k_y, k_z\,; n|\, \hat{x}\, |0, k_y, k_z\,; n\rangle\,. \tag{1.86}$$

This quantity is the so called *hybrid Wannier charge center (HWCC)*, and is close related to the Berry phase via the relation

$$\gamma_{_L} = \frac{2\pi}{a_x}\sum_n \bar{x}_n \tag{1.87}$$

where L is the path on which the hybrid Wannier orbitals were subjected to the Fourier transform. It is proved [84, 233] that the Chern number is written in terms of the HWCCs as

$$C = \frac{1}{a_x}\left[\sum_n \bar{x}_n(k_y = 2\pi) - \sum_n \bar{x}_n(k_y = 0)\right]. \tag{1.88}$$

In (1.88) the HWCCs $\bar{x}_n(k_y)$ are considered to be smooth functions of k_y, where $k_y \in [0, 2\pi]$. This condition is satisfied by constructing the HWCCs in the 1D maximally localized gauge[66] [231]. Then, the HWCCs are related to the eigenvalues of the Wilson loop with the relation [84, 252]

$$\bar{x}_i = \frac{a_x}{2\pi}\arg\lambda_i\,, \tag{1.89}$$

perhaps up to a reordering.

This close relation between the HWCCs and the Berry phase brings a physical interpretation of the Chern number. As the momentum (in the present case k_x, Fig. 1.14) varies in the Brillouin zone, the average position of the electrons, in the orthogonal direction, can change. Because of the periodicity of k_x, this average position must return to the unit cell, but also can ends up to another unit cell, different from the initial one. This represents a process of pumping charge, where in each cycle of k_x the charge moves by C unit cells, C being the Chern number.

[66] It is pointed out however, that the periodicity condition holds only modulo a lattice vector $R_x = na_x$, $n \in \mathbb{Z}$ [84].

1.5 Another view to the Chern number

The Chern number arose in §1.3.3 as the gauge-invariance modulo 2π of the Berry phase. But it can also be introduced in a more abstract mathematical way, independently of the Berry phase ideas. In this Section, the Chern number will be presented and examined briefly in the frame of Differential Geometry and Topology. This point of view is a glimpse to the deeper meaning of the Chern number, and enlights better its significance in mathematical physics.

Let M be a 2D compact manifold, and a map to a 2-sphere, $\boldsymbol{\varphi} : M \to S^2$.
This means that for $\mathbf{x} \in M$, it is $\boldsymbol{\varphi}(\mathbf{x}) = (\varphi_{x_1}(\mathbf{x}), \varphi_{x_2}(\mathbf{x}))$, with $\varphi_{x_1}^2 + \varphi_{x_2}^2 = 1$,
where $\mathbf{x} = (x_1, x_2)$ is a coordinate system on M.
On M can be defined the quantity $F_{ij} = \partial_i \varphi_j - \partial_j \varphi_i$,
and from this to define the topological invariant

$$C_1 = \frac{1}{2\pi} \int_M F_{x_1 x_2} \, dx_1 dx_2 \, . \tag{1.90}$$

This takes integer values, and it is just the Chern number. Although it seems innocence, (1.90) is remarkable in many respects. Firstly, F_{ij} looks a lot like a curvature tensor (hence the name "Berry curvature" in the Berry phase context). Furthermore, (1.90) reminds strongly the well known *Gauss-Bonnet theorem* [68, 174] which relates the Euler characteristic χ of a surface with its gaussian curvature K, namely[67]

$$\int_M K dA = 2\pi \chi(M) \, . \tag{1.91}$$

This theorem has quite a few generalizations, including the *Riemann-Roch theorem* (for Riemann surfaces) [174], and the *Atiyah-Singer index theorem* for differential operators on compact manifolds [174]. The version (1.90) or equivalently (1.91) examined here, is the 2D-case of the *Chern-Gauss-Bonnet theorem* that holds for even-dimensional, compact, boundaryless manifolds. The Euler characteristic is a topological invariant for compact connected surfaces, and can be used to identify a surface from a triangulation. The Chern number is simply one of the generalizations of the Euler characteristic encountered in the various versions of Gauss-Bonnet theorem.

It is emphasized that (1.90) holds for compact surfaces. For non-compact surfaces, (1.90) also holds if the curvature vanishes at large distances, so to converge the integral defining the Chern number. For convenience, it is usual to replace 2D non-compact surfaces by topologically equivalent[68] compact surfaces. To accomplish this, a process known in Topology as *compactification* can be applied [115]. With compactification, a non-compact space can be replaced by a compact one. There are two main ways of compactification; the most easy and common is the so called *one-point compactification* [115], in which all points at infinity are mapped to one point and this point is attached in the initial space, see Fig. 1.14a. For example, using one-point compactification, the

[67] Here for a closed orientable surface.

[68] In the sense of an homeomorphism, here denoted with $\sim$.

infinite 2D plane $M = \mathbb{R}^2$ is mapped with the stereographic projection to a sphere, in which the sphere's north pole corresponds to the points at infinity[69] [115]. In terms of Topology it is $M \sim S^2$, thus the initial mapping φ now can be considered to map spheres to spheres, $\varphi : S^2 \to S^2$.

In this context, the meaning of C_1 is that it counts how many times the first sphere wraps around the second, and its integer values characterize the second homotopy class, $\pi_2(S^2)$.

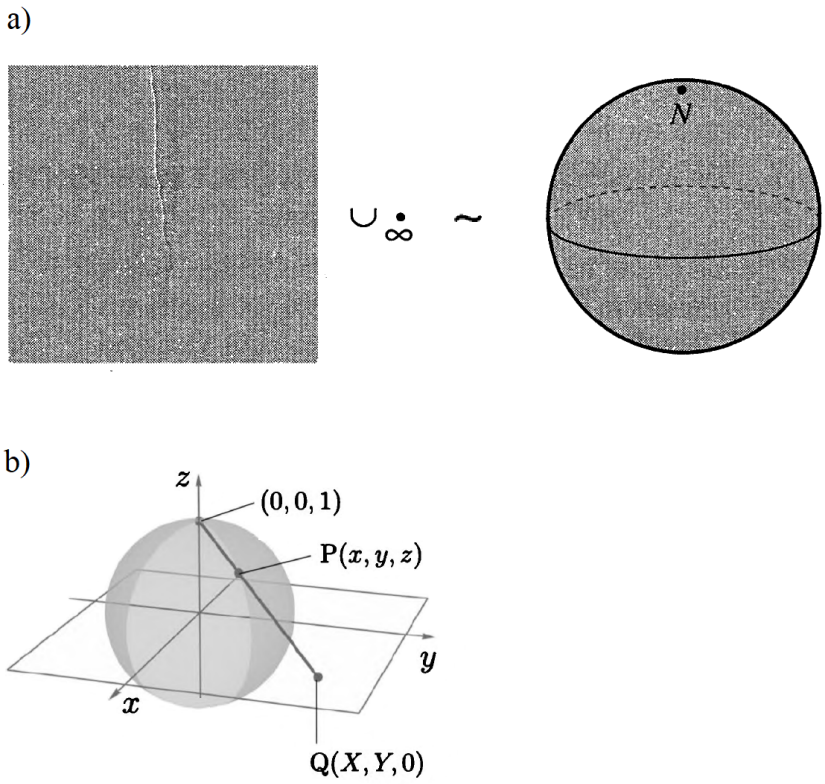


Figure 1.14: a) The infinite plane is non-compact; however, it can be replaced by a 2-sphere using one-point compactification. b) The stereographic projection is used for the mapping.

In physical applications, $\varphi(\mathbf{x})$ usually represents a point on the Bloch or Poincare sphere, x_1, x_2 can be wavevector (momentum) components, and the 2D manifold M can be a torus T^2 representing a Brillouin zone. Specifically for the case of a Brillouin zone, C_1 is often referred as the *TKNN invariant* [242], as already mentioned in §1.3.3. Another famous physical phenomenon of topological character, where C_1 appears inherently, is the quantum Hall effect.

C_1 can also defined to be the integral

$$C_1 = \int_M \omega \tag{1.92}$$

[69] In fact this is the Riemann sphere, known also from Complex Analysis or Differential Geometry.

where $\omega = 1/4\pi F_{\mu\nu}\, dx^{\mu} \wedge dx^{\nu}$ is a 2-form, known as the *first Chern class*. This integral is the first of an infinite sequence of Chern numbers defined in even number of dimensions. In specific, the nth Chern number is defined in $2n$ dimensions by an integral over the base of a $2n$-form with complex fiber, as

$$\begin{aligned} C_n &= \frac{1}{n!}\left(\frac{1}{4\pi}\right)^n \int e^{\mu_1\nu_1\mu_2\nu_2...\mu_n\nu_n} F_{\mu_1\nu_1} F_{\mu_2\nu_2} ... F_{\mu_n\nu_n}\, dx^{\mu_1} \wedge dx^{\nu_2} \wedge ... \wedge dx^{\nu_{2n}} \\ &= \frac{1}{n!} \int_M Tr(\omega^n) \end{aligned} \tag{1.93}$$

where the trace is over the degrees of freedom repesented by the fiber.
$e^{\mu_1\nu_1\mu_2\nu_2...\mu_n\nu_n}$ is the full antisymmetric tensor, vanishing if any of the indices are equal, equals $+1$ for an even permutation of $\{1, 2, ..., n\}$ and -1 for an odd permutation. The $2n$-form integrated is the so called nth Chern class. These higher Chern numbers and classes are not used in the theory of topological insulators but they are needed in other disciplines of physics, mainly in Quantum Field Theory.

2. Time Reversal Symmetry and introduction to Topological Insulators

2.1 Introduction

Topological insulators (TIs) are a type of electronic materials, recently discovered, which in the context of conductance classification exhibit a quite peculiar behavior. A topological insulator, unlike the ordinary insulators and metals, in the bulk of the material is insulating while at its surface is metallic. This unconventional character of electric conductance emanates from electronic states which due to topological reasons[1] appear at the surface of the material (or at the edges of the system if it has 2D geometry). These surface or edge states are protected from perturbations by time-reversal symmetry (TRS), and are characterized by nonzero Chern numbers. These uncommon properties set apart topological insulators from conventional insulators and metals.

Pure topological materials are not found in nature but they can be constructed using crystal synthesis techniques. The first TI discovered experimentally was a 2D quantum spin-Hall system, created by HgTe quantum-well structures. This system exhibits a Hall effect different from the classical one, specifically the quantum spin-Hall effect, which takes place in zero external magnetic field and can be observed through the conducting edge states forming a so called Kramers doublet. Other topological materials discovered later were strained 3D layers of HgTe, ternary tetradymite compounds (Bi_2Te_2S, Bi_2Te_2Se, Bi_2Se_2Te), the semiconducting alloy $Bi_{1-x}Sb_x$ and others. Many theoretical models of topological materials admit analytical solutions; this is very convenient as it helps to clearer illustrate important properties of the edge and surface states, like spin helicity and Dirac-like spectrum, and develops intuition for their behavior.

A very important property of topological materials is the protection of the helicity of the edge and surface states to the defects of the crystal (disorders, impurities etc) ˙ this makes them suitable for many intriguing applications such as spintronics, topological optical computing, topological lasers and others. This protection property stems from the aforementioned time-reversal symmetry – and in fact many important properties of TIs are close related to the preserving or breaking this symmetry.

In this chapter the basic concepts of 2D and 3D TIs are presented, mainly with intuitive arguments. At first, the time-reversal symmetry is examined, and its relation

[1] Specifically, due to the bulk-edge correspondence, see §1.4.3.

with the helical edge states[2] and the quantum Hall effect. Then, two types of topological systems are studied in detail, namely the Chern insulators and the topological insulators, in the 2D case; their main difference is that the quantum Hall effect without external magnetic field occurs in Chern insulators with the TRS broken, while in topological insulators occurs with the TRS preserved. These (not so simple) models help to enlight the topological origin of the helical edge states. It can be shown that the helical edge states stem from a boundary condition equivalent to a band-gap region barrier that describes the inversion of the band structure at the edges. Lastly, a short introduction to some models of 3D TIs and some other ancillary subjects are also presented.

2.2 Rudiments of Time Reversal Symmetry

Time-reversal symmetry (TRS) is the invariance of physical laws under time reversal transformation[3], namely

$$\widehat{T} : t \to -t \,. \tag{2.1}$$

TRS is a fundamental physical property, and for practical systems in many cases has vivid consequences. Systems exhibit quite different behavior depending on if they are TR-invariant or not, and both cases induce equally interesting physical phenomena[4].

A system is said to conserve the TRS, or to be TR-invariant, if its Hamiltonian commutes with the TR- operator[5],

$$[\widehat{H}, \widehat{T}] = 0 \,. \tag{2.2}$$

The application of $\widehat{T}$ to a particle with momentum **p**, spin **S** and position **r** changes these quantities as

$$\mathbf{r} \to \mathbf{r} \,, \tag{2.3a}$$

$$\mathbf{p} \to -\mathbf{p} \,, \tag{2.3b}$$

$$\mathbf{S} \to -\mathbf{S} \,, \tag{2.3c}$$

i.e., it reverses the direction of **p** and **S** at its current position.
The Hamiltonian of a TR-invariant system does not change under transformations (2.3). For a Hamiltonian that is an even function of momentum and is independent of spin, this holds trivially; e.g., a free non-relativistic particle. A more interesting case of TRS

[2] The meaning of helical states is explained later in this chapter.

[3] The terminology was first introduced by E. Wigner in 1932.

[4] For example, in systems with TRS breaking, reversing the motion of the particles in a magnetic field Hall voltages can occur. In contrary, in systems with TRS invariance, Hall effects cannot occur but other topological phenomena like the Z_2 topological classification take place.

[5] It is reminded from Quantum Mechanics that two operators that commute (here $\widehat{H}$ and $\widehat{T}$) are compatible (i.e., their observables can be measured simultaneously) and they have the same set of eigenvectors.

is a Hamiltonian which is odd in both momentum and spin; e.g., the case of a massless spin-$\frac{1}{2}$ particle moving with velocity v,

$$\widehat{H} = v\,\boldsymbol{\sigma}\cdot\mathbf{p} = \frac{2v}{\hbar}\,\mathbf{S}\cdot\mathbf{p}\,, \tag{2.4}$$

where the components of $\boldsymbol{\sigma} = (\sigma_x\,,\,\sigma_y\,,\,\sigma_z)$ are the Pauli matrices [178].

Most symmetries are represented in Quantum Mechanics by unitary operators; however, the operator $\widehat{T}$ describing TRS is *antiunitary*. An antiunitary operator[6], like a unitary one, preserves the norm of a vector $|\psi\rangle$ on which it acts, but is *antilinear*, i.e., for any scalar c it holds $\widehat{T}c|\psi\rangle = c^*\,\widehat{T}|\psi\rangle$. The representation of $\widehat{T}$ and the necessity for its antilinear property, can be found from the transformations (2.3) that must implies. From (2.3a,b), it is

$$\widehat{T}\,\mathbf{r}\,\widehat{T}^{-1} = \mathbf{r}\,, \tag{2.6a}$$

$$\widehat{T}\,\mathbf{p}\,\widehat{T}^{-1} = -\mathbf{p}\,. \tag{2.6b}$$

But it is also[7] $\widehat{T}\,[\mathbf{r},\,\mathbf{p}] = [\mathbf{r},\,\widehat{T}\,\mathbf{p}] = -[\mathbf{r},\,\mathbf{p}]$, therefore

$\widehat{T}\,[\mathbf{r},\,\mathbf{p}]\,\widehat{T}^{-1} = -[\mathbf{r},\,\mathbf{p}]\,\widehat{T}^{-1} = -i\hbar\,\widehat{T}^{-1}$ or

$\widehat{T}\,i\hbar\,\widehat{T}^{-1} = -i\hbar\,\widehat{T}^{-1}$ or eventually $\widehat{T}\,i = -i\,.$

Thus, the TR-operator must be proportional to the complex-conjugation operator $\widehat{K}$, and in general can be written as $\widehat{T} = \widehat{U}\widehat{K}$, where $\widehat{U}$ is unitary.
For spinless particles[8] this does not needed, and $\widehat{T}$ is just the complex-conjugation operator $\widehat{K}$,

$$\widehat{T}\psi(\mathbf{r}) = \widehat{K}\psi(\mathbf{r}) = \psi^*(\mathbf{r})\,, \tag{2.7}$$

and it is readily seen that $\widehat{T}^{\,2} = \widehat{1}\,.$
For spinfull particles, additional to (2.6a,b) must hold

$$\widehat{T}\,\mathbf{S}\,\widehat{T}^{-1} = -\mathbf{S}\,. \tag{2.6c}$$

---

[6] An operator is unitary when $\widehat{U}^{-1} = \widehat{U}^{\,\dagger}$, i.e., $\widehat{U}\widehat{U}^{\,\dagger} = \widehat{1}\,.$ It preserves the inner product,

$$\langle\bar{a}|\bar{b}\rangle = \langle a|\widehat{U}^{\,\dagger}\,\widehat{U}|b\rangle = \langle a|b\rangle\,.$$

An operator is antiunitary when $\widehat{U}^{-1} = -\widehat{U}^{\,\dagger}$, i.e., $\widehat{U}\widehat{U}^{\,\dagger} = -\widehat{1}\,.$ It preserves the inner product but additionally introduces a complex conjugation [210],

$$\langle\bar{a}|\bar{b}\rangle = \langle a|\widehat{U}^{\,\dagger}\,\widehat{U}|b\rangle = \langle a|b\rangle^* = \langle b|a\rangle\,. \tag{2.5}$$

[7] Using that $[\mathbf{r},\,\mathbf{p}] = i\hbar$ and the property of commutator $[A,\,BC] = B[A,\,C] + [A,\,B]C$, known from Quantum Mechanics [177]. In these relations, **p** and **r** in the commutators are considered to be operators.

[8] Although electrons are indeed spinors, there are cases where they are treated as spinless particles, e.g., in the Haldane model.

This action is realized by a rotation by π around some arbitrary axis – and the prevailing convention is to rotate spin around y-axis. $\widehat{T}$ must implement this rotation, and also must be proportional to the complex-conjugation operator $\widehat{K}$ as its action on **r** and **p**, (2.6a,b), must still hold regardless of whether the particle has spin or not. It can be shown [20, 210] that for spinors these requirements are satisfied setting the TR-operator as[9]

$$\widehat{T} = i\sigma_y \widehat{K}\,, \tag{2.8}$$

where $\sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}$ is the second Pauli matrix.

In this case, and in general for half-integer spin particles, it holds

$$\widehat{T}^{\,2} = -\widehat{1} \quad \text{and} \quad \widehat{T}^{\,-1} = -\widehat{T}\,. \tag{2.9a,b}$$

A very important consequence of the TR-invariance is the *Kramers theorem* :
in a TR-invariant system with half-integer spin particles[10], each energy level is at least two-fold degenerate (i.e., it belongs to at least two different eigenstates). This theorem is based on the property $\widehat{T}^{\,2} = -\widehat{1}$ and can be proved quite easily. Let consider two states $|u\rangle$, $|v\rangle$, related to each other by the TR-operator,

$$|v\rangle = \widehat{T}|u\rangle\,. \tag{2.10}$$

As $\widehat{T}$ commutes with the Hamiltonian,
$|u\rangle$ and $|v\rangle$ have the same energy eigenvalue λ because

$$\widehat{H}|v\rangle = \widehat{H}(\widehat{T}|u\rangle) = \widehat{T}\widehat{H}|u\rangle = \widehat{T}\lambda|u\rangle = \lambda(\widehat{T}|u\rangle) = \lambda|v\rangle\,,$$

using that λ is real.
Also, from (2.10) it is

$$\langle v| = \langle u|\widehat{T}^{\,\dagger} \tag{2.11}$$

and

$$\widehat{T}^{-1}|v\rangle = |u\rangle\,. \tag{2.12}$$

With all the above, and (2.5), the inner product of $|v\rangle$, $|u\rangle$ gives

$$\langle v|u\rangle = \langle u|\widehat{T}^{\,\dagger}\,\widehat{T}^{-1}|v\rangle = -\langle u|\widehat{T}^{\,\dagger}\,\widehat{T}|v\rangle = -\langle u|v\rangle^* = -\langle v|u\rangle\,, \tag{2.13}$$

from which follows that $\langle v|u\rangle = 0$ identically.
Therefore, $|u\rangle$ and $|v\rangle$ are orthogonal (i.e., independent and different), and have the same eigenvalue λ. This proves the theorem.

---

[9] The representation (2.8) holds only for spin-$\frac{1}{2}$ particles, not in general for spinfull particles.

[10] The system must be composed from an odd number of such particles. The reason has to do with the total angular momentum under TRS and will not be examined here.

Kramers theorem has interesting applications to electrons in crystals, where odd- and even-electron systems exhibit very different behavior. In crystalline systems, it can be proved [20] that TRS imposes a Bloch eigenfunction $|\psi_{n,\mathbf{k}}\rangle$ to be degenerate with a time-reversed corresponding one, according to

$$\widehat{T}|\psi_{n,\mathbf{k}}\rangle = e^{i\phi}\,|\psi_{n,-\mathbf{k}}\rangle\,, \tag{2.14}$$

where the phase ϕ is n and $\mathbf{k}$ dependent. This also holds for the cell-periodic Bloch functions $|u_{n,\mathbf{k}}\rangle$, while for the $\mathbf{k}$-dependent Hamiltonian[11] $\widehat{H}(\mathbf{k})$ it is

$$\widehat{T}\,\widehat{H}(\mathbf{k})\,\widehat{T}^{-1} = \widehat{H}(-\mathbf{k})\,. \tag{2.15}$$

At the points where $-\mathbf{k} = (\mathbf{k} \mod \mathbf{a})$, $\mathbf{a}$ being a reciprocal lattice vector, $-\mathbf{k}$ and $\mathbf{k}$ are simply duplicate labels and (2.15) becomes

$$\widehat{T}\,\widehat{H}(\mathbf{k})\,\widehat{T}^{-1} = \widehat{H}(\mathbf{k})\,. \tag{2.16}$$

At these special $\mathbf{k}$-points, the Kramers theorem is applied and gives that there all the states are doubly degenerate. These special wavevectors are known as *TR-Invariant Momenta (TRIM)* and are important for the topological states.

Even more profound consequences has the simultaneous presence of TR and the parity operator $\widehat{P}$. The combined operator $\widehat{P}\widehat{T}$ is an antiunitary operator which maps $\mathbf{k}$ to itself at all $\mathbf{k}$. In this case, the Kramers theorem implies that all bands are doubly degenerate, everywhere in the Brillouin zone. This is the expected from spin degeneracy in a nonmagnetic system in the absence of spin-orbit coupling; if inversion is also present, the Kramers theorem implies that the bands remain doubly degenerate even in the presence of spin-orbit coupling.

More details can be found in [20] and [252], and a more general discussion of the TR operator is in [210]. In the theory of topological insulators the Kramers theorem plays an important role as many of their properties can be interpreted with it.

2.3 Broken Time-Reversal Symmetry, Chiral Edge States and Quantum Hall Effect

Breaking the TRS can induce profound implications. In Fig. 2.1 it is shown an important case of TRS breaking. It is about a 2D electron gas[12] subject to a magnetic field $\mathbf{B}$ perpendicular to it. Under the influence of the magnetic field (Lorentz force), the electrons move on trajectories of two types. The first are closed circular orbits in the interior (bulk) of the system, away from its edges. As these orbits are localized, no net current flows in the central region; the current vanishes in the bulk.

[11] See F/note 55, p. 36.

[12] A 2D electron gas (2D-EG) is an electron gas that is free to move in two dimensions, but is strictly confined in the third. This strict confinement causes quantized energy levels for motion in the third direction, which in many problems can be ignored. Thus the electrons appear to form a 2D sheet.

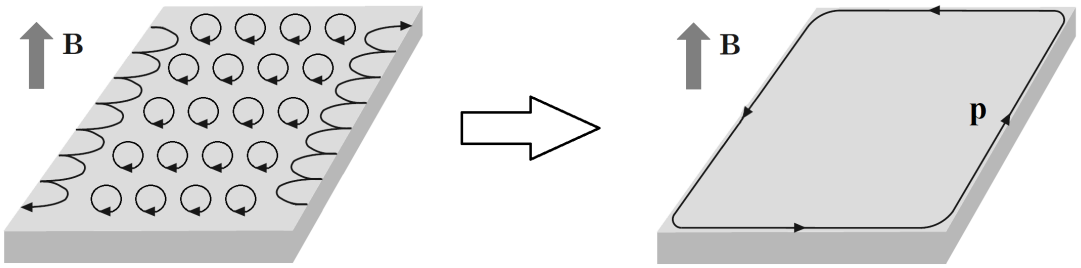


Figure 2.1: A 2D electron gas under a perpendicular magnetic field, and creation of edge states. (a) Classical intepretation. (b) Quantum interpretation.

The other type are the trajectories of the electrons at the boundary, which move skipping repeatedly along the edges. These orbits are open, skipping all the way around the boundary of the system; they constitute specific electronic states that carry the current and they are formed in a magnetic field. This edge current is unidirectional : it is strictly right-moving on one edge and left-moving on the opposite edge. The above arguments, Fig. 2.1a, although classical, is a physical interpretation correct in its essence.

When the magnetic field is strong enough, a full quantum treatment is necessary. In this case, the skipping motion of the carriers is quantized and the skipping trajectory becomes an 1D edge channel encircling the interior of the system [92, 162], Fig. 2.1b. In these edge states, the direction of the momentum **p** is strictly tied to the orientation of the magnetic field. This directionality the edge state has, protects it against perturbations of the material, which in every real system is always present, more or less. Besides the defects of the crystal (disorders, impurities), the main cause of perturbations[13] are the random fluctuations of the electrostatic potential of the background; these cause elastic scattering, flipping the momentum, $\mathbf{p} \rightarrow -\mathbf{p}$. Neverthless, for the edge state such backscattering events are strictly prohibited[14]. The strict directionality of the edge states is called *chirality*. Chirality is a manifestation of broken TRS that appears in the quantum Hall effect [129] and explains its dissipationless property.

What follows is a brief description of the Hall effect, giving emphasis in the formation of the edge states presented above. In the Hall effect, a conductor in which current I is flowing, is subject to an external static magnetic field **B**; then, a voltage V_H appears in direction transverse to the flowing current, Fig. 2.2. When the magnetic field is strong enough, the Hall conductivity $\sigma_H = I/V_H$ becomes quantized as[15]

$$\sigma_H = \frac{2e^2}{h} n\,, \tag{2.17}$$

[13] Especially at low temperatures.

[14] Except only if there is also a counter-directional state in the same edge.
It is also considered that the edge states between opposite edges cannot interact by quantum tunneling.

[15] So this is about the (integer) quantum Hall effect [129].

i.e., in multiples of e^2/h, where e is the electron charge, $n = 1, 2, \ldots$, and the factor 2 in (2.17) stems from spin degeneracy.

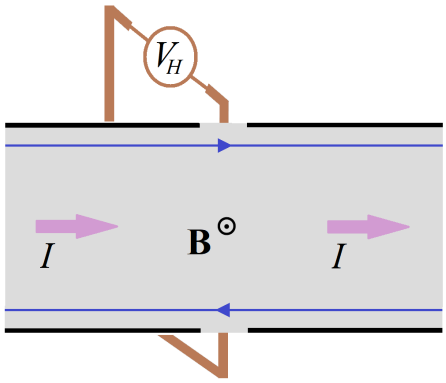


Figure 2.2: The geometry of Hall effect, with the edge states shown at the edges of the system. The edge states are chiral.

In such a setup, the longitudinal conductivity (i.e., in the direction of the flowing current) theoretically is zero; in reality it is very small, practically zero. In this quantum status the electric current does not prodcuce resistive loses. This dissipationless property is due to the chiral edge states, which carry the electric charge without scattering – hence without resistance along the edge. The necessary condition for this regime to appear, is that bulk electric carriers must not exist. Consequently, in a quantum Hall system, the edge states must appear in the energy gap separating the bulk bands, Fig. 2.3.
Furthermore, if the band structure is continuously deformed but leaving intact the gap between the bulk bands, then the quantum Hall effect (and the edge states) remains unaffected, Fig. 2.3b,c ˙ this means that the edge states are robust in such deformations.

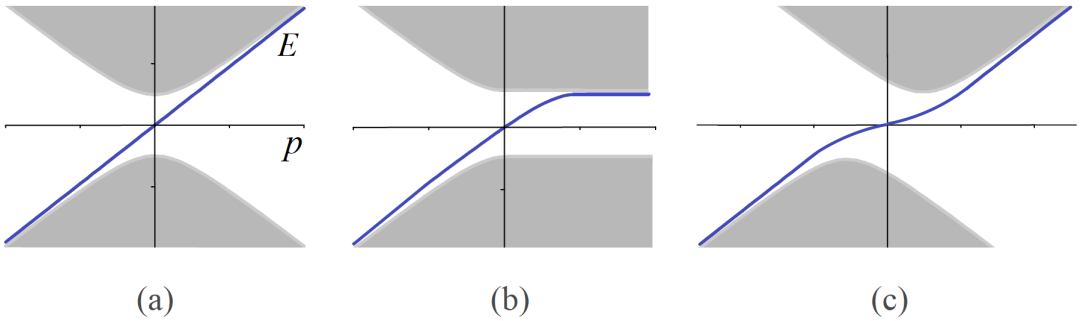


Figure 2.3: A schematic example of band structure of a quantum Hall system. The gray cones are the bulk states, the blue line is the edge state. (a) Original case. (b), (c) Cases arising by deforming continuously the (a), without touching the gap and the edge state.

2.4 Helical states and the concept of the 2D Topological Insulators

Besides the quantum Hall effect, the robust edge states is the keystone in the physics of the 2D topological insulators[16] (2D-TIs). As discussed above, in the Hall effect the edge states appear when there is a magnetic field quite strong to break the TRS and quantize the Hall conductivity. In contrary, in 2D-TIs the edge states do exist without a magnetic field, i.e., without breaking the TRS[17]. This is feasible because of the spin-orbit coupling in the TI materials. The spin-orbit coupling, in its simplest model for a spin-$\frac{1}{2}$ material, can be considered as an intrinsic effective magnetic field $\mathbf{B}_{eff}$, pointing at opposite directions for the up- and down-spin values. Concerning the 2D-TIs, each case (spin-up, spin-down) can be considered as a copy of a quantum Hall insulator, having a gapless edge state with opposite propagation and field directions in each one, Fig. 2.4. These two subsystems together result in a pair of edge states in a zero magnetic field – and this is a simple model for a 2D-TI. In 2D-TIs, the defining feature of the edge states is the locking between the directions of spin and momentum, known as *helicity*. Two such edge states constitute a *Kramers doublet*; they have linear dispersion and cross each other in the gap of the bulk band, at a so called *Dirac point*, Fig. 2.5.

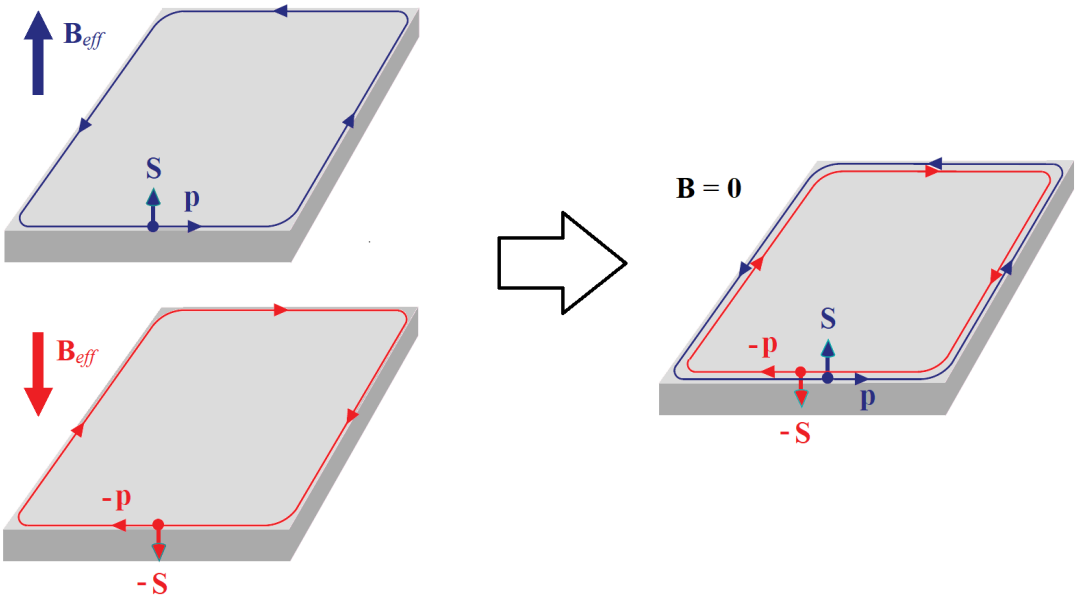


Figure 2.4: (adapted from [245]). Schematic model of a 2D-TI as a superposition of two quantum Hall systems. The edge states are helical.

[16] See [116, 117, 18, 19, 133, 134, 207].

[17] But if a 2D-TI is subject to a magnetic field strong enough, then the TRS breaks in it too.

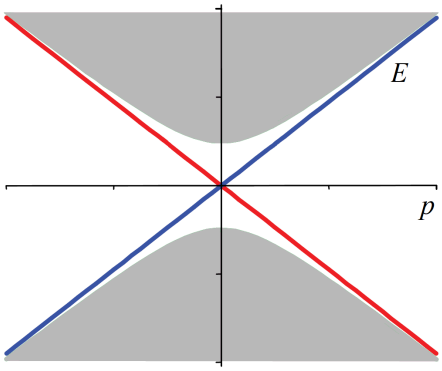


Figure 2.5: Schematic example of band structure of a 2D-TI.
The upper cone corresponds to the conduction band, the bottom to the valence band. The two helical edge states, corresponding to the up- and down-spin, are indicated.

In 2D-TIs the edge states do exist for both propagating directions; however, for scattering potentials of single particles that preserve the TRS, the scattering between these paired edge states is impossible. A simple explanation is the following. The paired helical edge states[18] $|R\rangle$, $|L\rangle$, can be transformed one to the other by reversing both the momentnum and spin, Eqs. (2.3b,c). This means the time is reversed, Eq. (2.1), hence the two edge states are related by the TR-operator as $|L\rangle = \widehat{T}|R\rangle$. If there is no scattering between the edge states, then their interaction via the scattering potential V must be null; in matrix representation this is

$$\langle L|\widehat{V}|R\rangle = 0\,. \tag{2.18}$$

It will be proved that (2.18) indeed holds.
As the potential V is TR-invariant it satisfies that $\widehat{T}\,\widehat{V}\,\widehat{T}^{-1} = \widehat{V}$ or

$$\widehat{T}\,\widehat{V} = \widehat{V}\,\widehat{T}\,. \tag{2.19}$$

Also, from $|L\rangle = \widehat{T}|R\rangle$ it is

$$\langle L| = \langle R|\widehat{T}^{\dagger}, \tag{2.20}$$

and

$$\widehat{T}^{-1}|L\rangle = |R\rangle\,, \tag{2.21}$$

while it is reminded that for $\widehat{T}$ holds

$$\widehat{T}^{\,-1} = -\widehat{T}^{\dagger} = -\widehat{T}\,. \tag{2.22}$$

[18] where R, L denotes the right- and left-moving respectively.

Using (2.5) and all the above, it can be written

$$\begin{aligned}\langle L|\widehat{V}|R\rangle &= \langle R|\widehat{T}^{\dagger}\,\widehat{V}\,\widehat{T}^{-1}|L\rangle = -\langle R|\widehat{T}^{\dagger}\,\widehat{V}\,\widehat{T}|L\rangle = -\langle R|\widehat{T}^{\dagger}\,\widehat{T}\,\widehat{V}|L\rangle \\ &= -\langle R|\widehat{T}^{\dagger}\,\widehat{T}\,\big(\widehat{V}|L\rangle\big) = -\langle R|\widehat{V}|L\rangle^{*} = -\langle L|\widehat{V}^{\dagger}|R\rangle \\ &= -\langle L|\widehat{V}|R\rangle\,, \end{aligned} \tag{2.23}$$

where $\widehat{V}|L\rangle$ was considered as an intermediate auxiliary ket, and $\widehat{V}^{\dagger} = \widehat{V}$ because $\widehat{V}$ is taken to be hermitian. From (2.23) it is deduced that (2.18) is true.

The resultant of (2.18) is that the perturbations of the potential background in the material cannot reverse the propagation direction of the helical edge states. The obstacles in the background can modify more or less the trajectories of the edge states; however, the two edge states always remain a "time-reversed pair" satisfying the Kramers theorem, and their conduction capability in the gobal sense is the same as in the ideal case. This is a manifestation of the key property of topological insulators to preserve invariant their global characteristics under continuous local deformations[19].

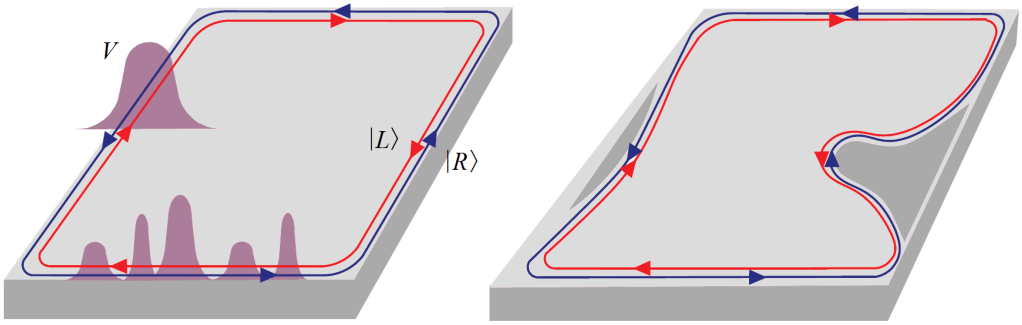


Figure 2.6: (adapted from [245]). The robustness of the helical edge states to the perturbations of the background potential and material defects.

The topological robustness of the helical edge states can be better emphasized by comparing them with the ordinary 1D conductors in presence of potential perturbations. In the ordinary conductors, the random perturbations of the potential cause backscattering; the result is all the propagating states to become localized, a phenomenon known as *Anderson localization* [6]. This takes place no matter how weak the perturbation or the disorder is, as long as the system is large enough. But for TIs, because of (2.18), Anderson localization is impossible, and this is a crucial qualitative difference from the ordinary conductors.

Due to the spin-momentum locking (aka helicity) in the edges, the electronic state of the 2D-TIs is also called "quantum spin-Hall state". It was theoretically predicted for graphene with spin-orbit coupling [116, 117], and also for semiconductor quantum wells

[19] Other topological characterists of TIs is the band inversion and the misc topological numbers [245].

[18, 19]. The quantum spin-Hall state was observed and investigated experimentally for first time in HgTe quantum well structures [133, 134, 207].

2.5 The concept of 3D Topological Insulators

The setup and physics of 2D-TIs have a corresponding case in 3D space : it is about the 3D topological insulators[20] (3D-TIs). In the surface of a 3D-TI robust electronic states are formed, topologically protected, while the bulk is insulating. The surface carriers can move freely in two dimensions; however, as with edge states, their spin is locked in the direction of the momentum, Fig 2.7.

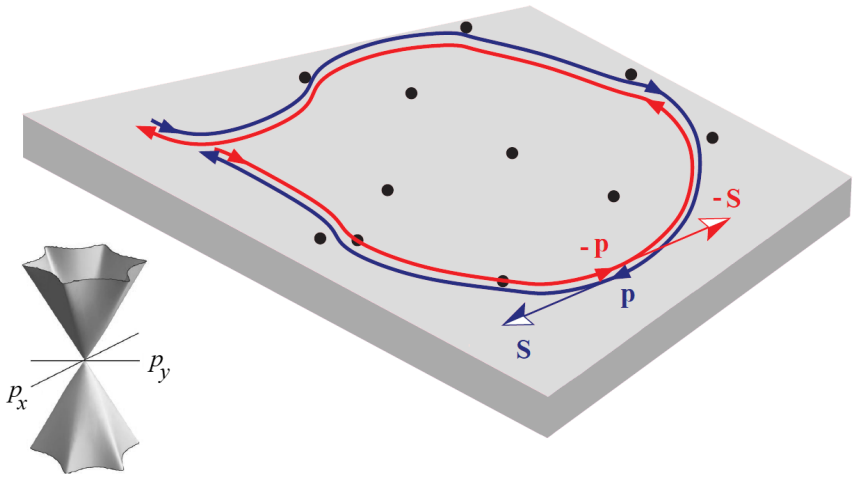


Figure 2.7: (adapted from [245]). Propagating states on the surface of a 3D topological insulator, and their dispersion diagram.

The simplest Hamiltonian able to describe surface states with such spin-momentum locking has already mentioned in (2.4), namely

$$\widehat{H} = v\,\boldsymbol{\sigma}\cdot\mathbf{p} = \frac{2v}{\hbar}\,\mathbf{S}\cdot\mathbf{p}\,. \tag{2.4}$$

In this case, the momentum is a vector on the surface, $\mathbf{p} = (p_x, p_y, 0)$.
In matrix form this Hamiltonian is written as

$$\widehat{H} = v\,(\sigma_x p_x + \sigma_y p_y) = v\begin{bmatrix} 0 & p_x - ip_y \\ p_x + ip_y & 0 \end{bmatrix}. \tag{2.24}$$

[20] See [95, 198, 69, 70, 173, 99, 100].

The energy $E_{\mathbf{p}}$ of a surface state with momentum $\mathbf{p}$ is the solution of the eigenvalue equation

$$\widehat{H}\psi = v\,\boldsymbol{\sigma}\cdot\mathbf{p}\,\psi = E_{\mathbf{p}}\,\psi\,, \tag{2.25}$$

where the wavefunction ψ here is a two-component spinor, $\psi = \begin{bmatrix}\psi_\uparrow \\ \psi_\downarrow\end{bmatrix}$.
The solution of (2.25) is

$$E_{\mathbf{p}} = \pm v\,|\mathbf{p}| = \pm v\sqrt{p_x^2 + p_y^2}\,. \tag{2.26}$$

Eq. (2.26) is indeed the dispersion relation for the 3D-TI (for the Hamiltonian (2.4), adopted in this case). Its graph is a double cone, where the up-cone ($+$ branch of $E_{\mathbf{p}}$) corresponds to the conduction band, and the down-cone ($-$ branch of $E_{\mathbf{p}}$) corresponds to the valence band of the surface carriers. This conical energy spectrum reminds strongly that of an ultra-relativistic electron ˙ due to this similarity, sometimes the surface states are called “Dirac fermions”.

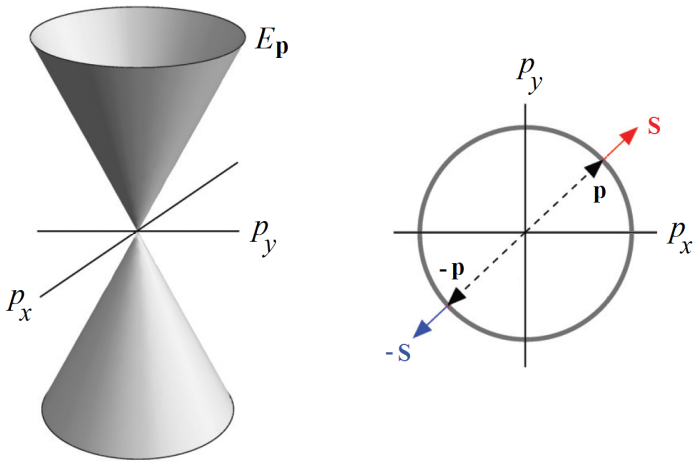


Figure 2.8: The energy spectrum (2.26), and a pair of surface states on a level of contant energy. The spin-momentum locking is also indicated.

A quite large number of materials and heterostructures support surface states, thus being 3D-TIs. For the first time the surface states predicted theoretically at interfaces between normal and inverted semiconductors (e.g., [257, 189]). Subsequently, quite a few materials found to be 3D-TI candidates, the most important of which are :

- the semiconductor alloy $Bi_{1-x}Sb_x$ [69],
- strained layers of Sn and HgTe [69],
- tetradymite compounds Bi_2Se_3, Bi_2Te_3 and Sb_2Te_3 [280],
- ternary chalcogenites based on Tl (Thalium) : $TlBiTe_2$, $TlBiSe_2$ [272, 152, 58],

- layered chalcogenites based on Pb [59, 109].

Indeed, the existence of topological surface states was experimentally verified and investigated for most of the above materials :

- in the semiconductor alloy $Bi_{1-x}Sb_x$ [99],
- in Bi_2Se_3 [270] and Bi_2Te_3 [44],
- in $TlBiTe_2$ [45] and $TlBiSe_2$ [212, 137, 45]
- in strained layers of HgTe [38],
- in $Pb(Bi_{1-x}Sb_x)_2Te_4$ [235] and $PbBi_2Te_4$ [139].

It is emphasized that the Dirac cone described by (2.26) is an ideal case, and for the simplest Hamiltonian for 3D-TIs. In reality, away from the Dirac point ($\mathbf{p} = \mathbf{0}$), higher order terms of $\mathbf{p}$ enter in (2.26) distorting the form of the double cone. These terms destroy the symmetry of positive and negative branch of $E_{\mathbf{p}}$ (aka *particle-hole symmetry*), hence the symmetry between up and down cones, Fig. 2.9a. Moreover, in tetradymite compounds Bi_2Se_3, Bi_2Te_3 and Sb_2Te_3 *hexagonal warping* effects take place [71, 153, 3, 138], causing anisotropy in the surface states, Fig. 2.9b. Another interesting variation appears in thin films of TI materials, where the topology of the bands is qualitatively different; instead of the gapless Dirac-like cone, a semiconductor-like spectrum with an energy gap at $\mathbf{p} = \mathbf{0}$ is formed [150, 122], Fig. 2.9c. Nevertheless, despite these variations, the notifying property of 3D-TIs is the Dirac-like dispersion, and their basic model is the Hamiltonian (2.24) as it causes a pair states, topologically protected and related to each other by the TR-operator.

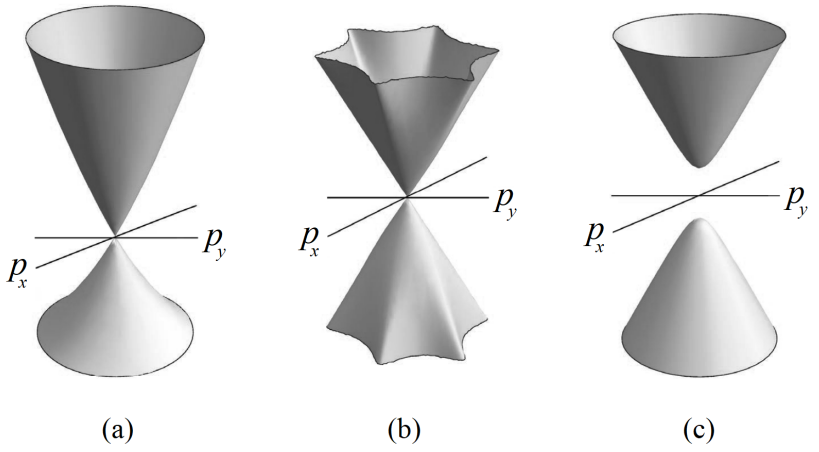


Figure 2.9: Variations in energy spectrum graph of surface topological states. (a) Broken particle-hole symmetry, (b) hexagonal warping, (c) energy gap.

2.6 Appendix : The Bulk-Edge Correspondence in Topological Photonic Structures[21]

The bulk-edge correspondence, already discussed in §1.4.3, relates the Chern topological numbers with the net number of unidirectional states supported at an interface of the relevant materials. This fundamental principle is perhaps the most consequential result of topological photonics, as it determines the precise physical manifestations of nontrivial topological features. Even though the bulk-edge correspondence has been extensively discussed and used in the literature, it seems that in the general photonic case with dispersive materials it has no solid mathematical foundation and is essentially a conjecture. In this section it is presented rigorous physically-motivated demonstration of this fundamental principle by showing that the thermal fluctuation-induced light-angular momentum spectral density in a closed cavity can be expressed in terms of the photonic gap Chern number, as well as in terms of the net number of unidirectional edge states. In particular, it is highlighted the rather fundamental connections between topological numbers in Chern-type photonic insulators and the fluctuation-induced light momentum.

2.6.1 Proof of the Bulk-Edge Correspondence in topological photonic structures

It can be proved [225] that in a topological system the gap Chern number is linked to the net number of unidirectional edge states as

$$C = -\sum_{\omega_m=\omega} s_m \,. \tag{2.27}$$

Thus, the Chern number of the bulk region determines precisely the net number of edge modes circulating around the lateral "opaque-type" walls of the closed cavity. In particular, a nontrivial Chern number implies the emergence of unidirectional gapless edge modes. For a positive (negative) gap Chern number the unidirectional modes propagate clockwise (anticlockwise) with respect to the z axis.

This result may be further generalized to give the number of edge modes propagating at the interface of two topological materials : the bulk-edge correspondence. To this end, let be considered the geometry depicted in Fig. 2.10, which shows a cavity half filled with two photonic insulators (the two materials share a photonic band gap). The cavity lateral walls are assumed opaque. Let C_1 and C_2 be the gap Chern numbers for material 1 and 2, respectively. Eq. (2.27) implies that C_1 and C_2 determine the number of modes propagating in the clockwise direction around the cavity walls, see Fig. 2.10. Hence, the number of modes propagating at the interface of the two materials (along the $+x_1$ direction) must be precisely $C_2 - C_1$, i.e., the gap Chern number difference.

[21] In this appendix it is presented the bulk-edge correspondence in the context for Topological Photonics; it is reproduced from [225].

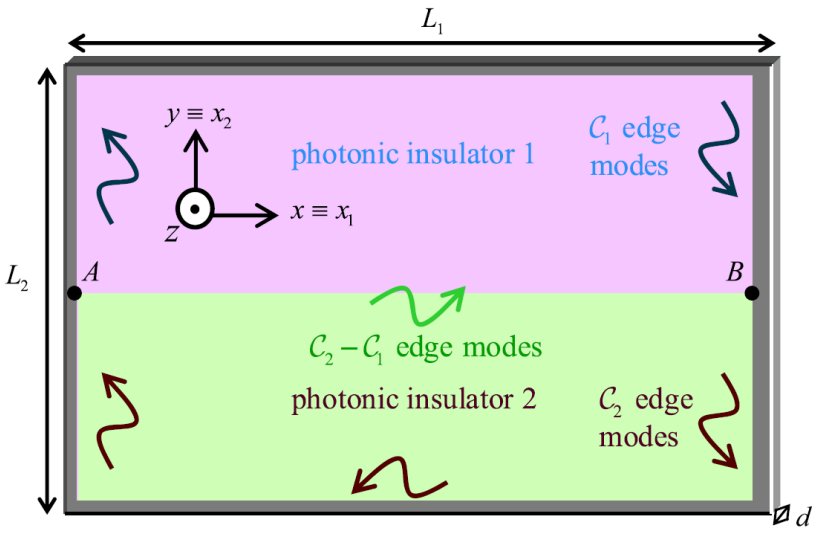


Figure 2.10: (reprinted from [225]). Illustration of the bulk-edge correspondence principle. A cavity (terminated with "opaque-type" lateral walls) is filled with two photonic insulators. The points A and B represent the two junctions.

The reason why this needs to be so is that otherwise the system would be unstable and a steady state could not be reached. Indeed, suppose that the net number of unidirectional modes propagating at the interface of the two materials is different from $C_2 - C_1$. In this situation, for one of the junction points (let us say point B in Fig. 2.10) the number of edge modes arriving at the junction is larger than the number of edge modes propagating away from the junction. Since the system response is linear, this implies that it would be possible to choose the complex amplitudes of the incident waves in such a way that the edge waves propagating away from the junction are not excited. But then, since by assumption there is no loss and there are not scattering channels available, the energy incident in the junction must remain stored in it. Hence, it is impossible to reach a stationary state for a time-harmonic excitation : the energy stored at the junction grows linearly with time similar to a lossless LC circuit excited at the resonance. Physically this is not acceptable, and hence the net number of unidirectional edge modes propagating at the interface of the two materials must be precisely $C_2 - C_1$. This concludes the proof of the bulk-edge correspondence principle.

To illustrate the application of the developed theory, let consider a 2D photonic crystal (the condition $\partial/\partial z = 0$ is enforced) formed by square-shaped nonreciprocal inclusions organized in a square lattice with period a, Fig. 2.11. The inclusions stand in air and are spaced by d. Furthermore, the analysis is restricted to TM polarized waves with nontrivial field components H_z, E_x, E_y. The electric response of the inclusion is assumed to be gyrotropic with the same dispersion model as a lossless magnetized plasma [26] (e.g., a magnetized semiconductor [187]), $\bar{\varepsilon} = \varepsilon_t \mathbf{1}_t + \varepsilon_a \hat{\mathbf{z}} \otimes \hat{\mathbf{z}} + i\varepsilon_g \hat{\mathbf{z}} \times \mathbf{1}$ where

$$\varepsilon_t = 1 - \frac{\omega_p^2}{\omega^2 - \omega_c^2}, \qquad \varepsilon_t = 1 - \frac{\omega_p^2}{\omega^2}, \qquad \varepsilon_t = \frac{1}{\omega} \frac{\omega_c \, \omega_p^2}{\omega_c^2 - \omega^2}, \tag{2.28}$$

and $\mathbf{1}_t = \hat{\mathbf{x}} \otimes \hat{\mathbf{x}} + \hat{\mathbf{y}} \otimes \hat{\mathbf{y}}$. In the above, ω_p is the plasma frequency, $\omega_c = -qB_0/m^*$ is the cyclotron frequency (positive when the magnetic field is oriented along $+z$), $q = -e$ is the electron charge, and m^* is the electron effective mass [26].

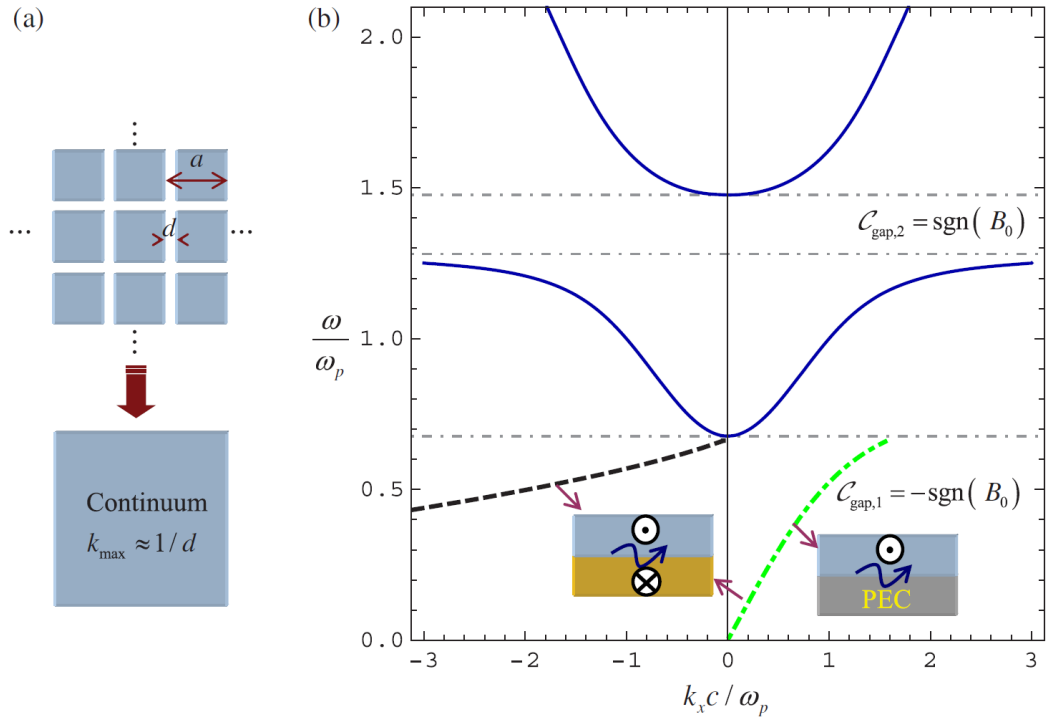


Figure 2.11: (reprinted from [225]).
(a) Geometry of a 2D photonic crystal formed by square-shaped gyrotropic-material inclusions organized in a square lattice. For sufficiently low frequencies the photonic crystal may be regarded as a continuum with a spatial frequency cutoff k_{max}.
(b) Band structure of material (solid blue lines) and dispersion of the edge states in the first band gap for (i) a gyrotropic-PEC interface (dot-dashed green line) and (ii) gyrotropic-gyrotropic interface with the materials biased with magnetic fields oriented in opposite directions (dot-dashed green line and dashed black line). The band structure, the edge-state dispersions, and the gap Chern numbers (indicated in the insets) are found using the continuum approximation.

The structural parameters of the photonic crystal are $a = (2\pi/5)(c/\omega_p)$ and $d = 0.1a$. For $\omega \ll \omega_p$ the air gaps are deeply subwavelength, and hence, in the long-wavelength limit it seems reasonable to approximate the photonic crystal by a continuum with the same permittivity as the inclusions, as illustrated in Fig. 2.11a. This approximation greatly simplifies the calculation of the band structure and of the gap Chern numbers. In order that the electromagnetic continuum is topological, it is necessary to impose a high frequency spatial cutoff k_{max} [222]. For the physical reasons discussed in detail in [223] the spatial cutoff should be taken on the order of $k_{max} \approx 1/d$. The photonic band

structure obtained with the continuum approximation is depicted in Fig. 2.11b (solid blue lines) for $\omega_c = \pm 0.8\omega_p$. As shown, there are two band gaps and the corresponding gap Chern numbers are indicated in the insets. The Chern number calculation is done as in [224] and takes into account the contribution of the negative frequency bands (not shown in Fig. 2.11b).

Next, let consider the low-frequency band gap for which the continuum approximation is more accurate. Its gap Chern number is $C_{gap,1} = -\,\mathrm{sgn}(B_0) = -\,\mathrm{sgn}(\omega_c)$, and thus it is topologically nontrivial. Hence, if the material is paired with a PEC boundary, the bulk-edge correspondence predicts that there is a single edge state propagating along the $+x$ direction. To confirm this prediction, it is used the continuum approximation to compute the edge state's dispersion. The spatial cutoff k_{max} is taken into account using the spatially dispersive model described in [223]. The calculated dispersion (for a material biased with $B_0 > 0$ and $\omega_c = 0.8\omega_p$). is plotted with a green dotted line in Fig. 2.11b, and yields the unidirectional gapless edge mode.

It is also interesting to analyze the case in which two topologically distinct plasmas are paired to form an interface inset of (Fig. 2.11b). In this scenario, the top region ($y > 0$) is biased with $B_0 > 0$ ($\omega_c = +0.8\omega_p$) and the bottom region ($y < 0$) is biased with $B_0 < 0$ ($\omega_c = -0.8\omega_p$). In this case, the gap Chern number difference is $-1 - 1 = -2$, and hence the bulk-edge correspondence predicts two modes propagating along the $+x$ direction. This property is confirmed by the numerical results : the edge-state dispersion is now formed by two branches. Because of the symmetry of the structure, one of the branches (with $k_x > 0$) is coincident with the one obtained for the gyrotropic-PEC interface geometry discussed previously. The second branch has $k_x < 0$ but a positive group velocity; i.e., it is a backward wave. Thus, in agreement with the bulk-edge correspondence, both edge modes propagate along the $+x$ direction.

To further validate the analysis and the link between the angular momentum and the gap Chern numbers, it was used the software CST MICROWAVE STUDIO to simulate the full wave response of a photonic crystal cavity with a geometry analogous to that of Fig. 2.10. The cavity lateral walls are PEC. The top region ($y > 0$) is a truncated photonic crystal with $\omega_c = +0.8\omega_p$, and the bottom region ($y < 0$) is a truncated photonic crystal with $\omega_c = -0.8\omega_p$. The structural parameters of the photonic crystals are as in the previous example. The CST simulations fully take into account the granular structure of the photonic crystals (the continuum approximation is not used). From the continuum results (Fig. 2.11), one may expect that for low frequencies this system supports (i) one unidirectional edge state propagating along the lateral walls, and (ii) two distinct unidirectional edge state,s propagating along the interface ($y = 0$) of the two gyrotropic photonic crystals. To test these ideas, the cavity was excited with a dipole-type antenna placed in between the two photonic crystals near to the left-hand side lateral wall. Figs. 2.12a and 2.13a show a time snapshot of the excited magnetic field (H_z) for a dipole oriented perpendicular (vertical dipole) and parallel (horizontal dipole) to the interface, respectively, with oscillation frequency $\omega = 0.5\omega_p$. The effect of weak material loss is taken into account to ensure the convergence of the simulations. The propagation of edge states at the lateral walls and at the interface of the two photonic crystals is evident. Furthermore, as can be seen from the Poynting vector lines in Figs. 2.12b and 2.13b, the energy circulates in closed orbits, such that for the top region (with gap Chern number $C_{gap,1} = -1$), the energy flows in the anticlockwise direction whereas in the bottom region (with gap Chern number $C_{gap,1} = +1$) it flows in the clockwise direction.

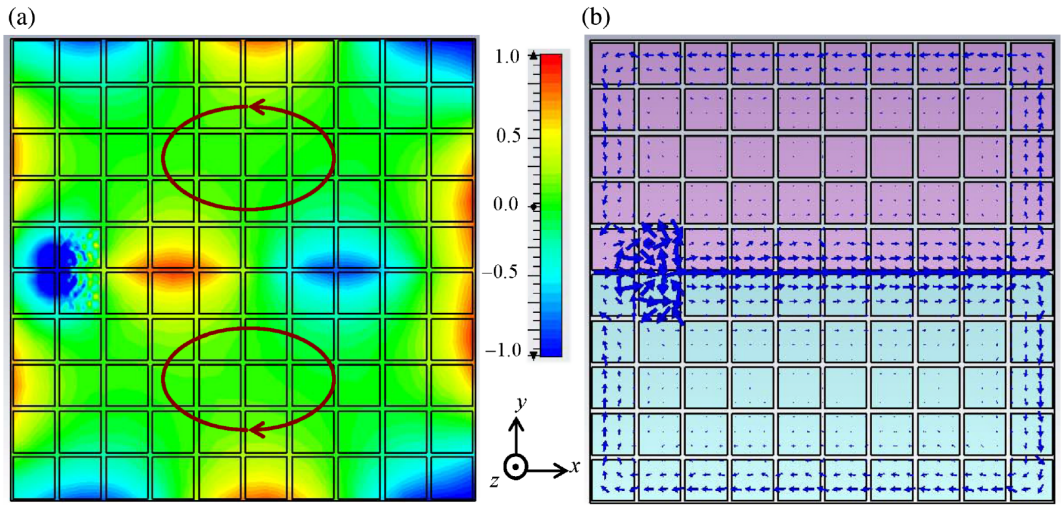


Figure 2.12: (reprinted from [225]). Photonic crystal cavity terminated with PEC lateral walls. The region $y > 0$ (top half of the cavity) is biased with a positive (along $+z$) magnetic field ($\omega_c = +0.8\omega_p$), and the region $y < 0$ (bottom half of the cavity) with a negative (along $-z$) magnetic field ($\omega_c = -0.8\omega_p$). The cavity is excited with a vertical (along $+y$) short electric dipole placed at the interface of the two regions near the left-hand side lateral wall. The oscillation frequency of the dipole is $\omega = 0.5\omega_p$.
(a) Time snapshot of the magnetic field H_z.
(b) Poynting vector lines showing how the energy circulates in the cavity.

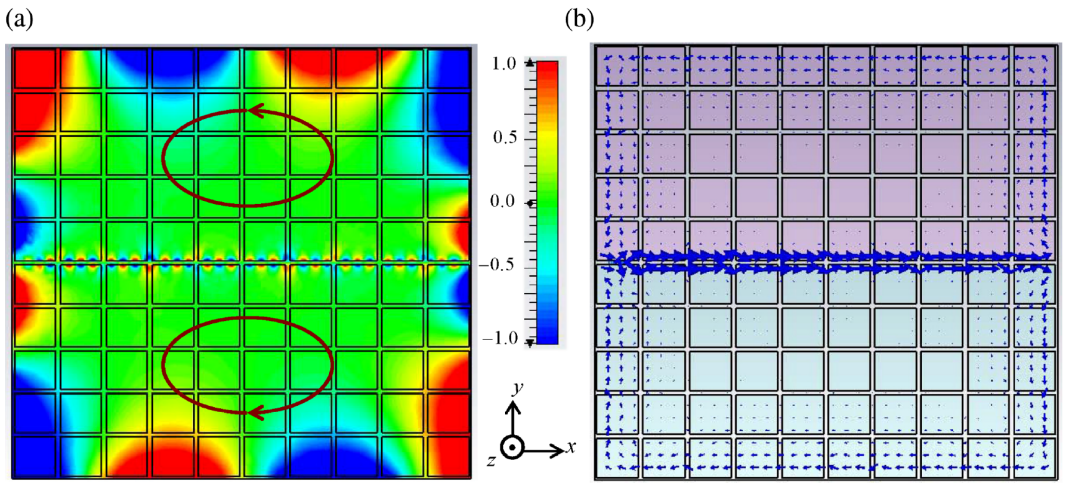


Figure 2.13: (reprinted from [225]). Similar to Fig. 2.12 but for a horizontal (along $+x$) electric dipole.

This result is in agreement with Eq. (2.27), which links the sign of the Chern number with the direction along which the energy circulates. Interestingly, the time animations available in the Supplemental Material[22] show that the edge state excited at the interface of the two photonic crystals ($y = 0$) by the vertical dipole is a forward wave (Fig. 2.12), whereas the edge state excited by the horizontal dipole is a backward wave (Fig. 2.13). Hence, in agreement with the dispersion of the edge states obtained with the continuum approximation (Fig. 2.11b), the interface $y = 0$ supports two unidirectional edge modes : a forward wave and a backward wave. Furthermore, as seen in Figs. 2.12 and 2.13, H_z has even (odd) symmetry with respect to $y = 0$ for the forward (backward) mode, respectively. The edgemode profiles obtained with the continuum theory have the same symmetries, which further reinforces the validity of this approximation.

[22] The link for the time animations of Figs. 2.12a and 2.13a is http://link.aps.org/supplemental/10.1103/PhysRevX.9.011037

3. Elements and applications of Topological Photonics

3.1 Introduction

It has been shown recently that many topological phenomena concerning the electronic band structure in solids, have analogues in optical and photonic systems too [161, 121, 126, 184, 89, 214]. The beginning was done by Haldane and Raghu who suggested the creation of a model of the anomalous quantum Hall effect in optical systems [91, 200]. Since then, it was found that optical systems can support unidirectional topological edge states, gapped photonic bands with nonzero Chern number, and photonic models of the quantum spin Hall effect [87, 88]. Based on these models, a plethora of applications was proposed, some very interesting ones, which will be discussed briefly in this chapter.

Photonic systems differ significantly from the electronic ones, and pose particular difficulties that must to be overcome. The main such difficulties are the following :

- The Hall effect and many other effects of topological nature require breaking of time reversal symmetry. But photons have no charge, hence this cannot be done trivially using magnetic fields as in electronic systems.
- The photons are bosons and tend to cluster to the lowest available energy level, instead of separating to distinct, well-separated bands.
- In contrast to electrons, photons do not interact directly with each other[1].
- Photons propagate at the speed of light, consequently the manipulation of their properties must either be done very quickly or be distributed spatially over the propagation path.
- Photons tend to be absorbed or scattered out of the system, hence a constant influx of new photons in the system is required.

In some cases the last point is advantageous. In an optical system, loss and gain are non-hermitian processes, and it has been found that non-hermitian processes offer new capabilities for topological phenomena[2], among them topological lasers presented below.

Besides the aforementioned difficulties, photonic systems have some significant advantages over the electronic ones. For example, changing the system parameters in

[1] However they can be forced to interact indirectly, via nonlinear interactions through a crystal lattice.

[2] See [279, 159, 57, 208, 145, 149, 217, 81, 269].

an optical system is generally easier than in a conventional solid-state system. Even more important, to observe topological phenomena in optical systems does not require extreme cooling, as is usually necessary in solid-state and atomic systems.

From the discussion in §2.3 it is obvious that the existence of quantum Hall effect depends on breaking the time reversal symmetry, by applying a magnetic field. Also, it was found that the the existence of topological phases depends on the presence or not of charge conjugation symmetry and of chiral symmetry[3]. Taken account how all these symmetries define the topological phases, a classification of the topological phases can be done, which sometimes refered as "the periodic table" of topological insulators [125, 229].

To achieve the symmetry conditions which allow the existence of topological phases in Photonics, a variety of techniques can be used. For example :

- to break the TRS, the so called Faraday rotators can be used [130],
- to create a chiral symmetry, the light can be directed to two-part optical systems in which two distinct types of optical components alternate,
- the system can be designed to have some form of periodic driving,

etc. These techniques allow the realization of topological states in quite a few systems ˙ e.g., in coupled resonant oscillator systems, in cold atom optical lattices, and in photonic quantum walk systems.

In this chapter will be discussed the topological effects occurring in waveguides, optical resonators, and dielectric photonic crystals. Simple quantum walk systems for light are discussed as well. The branch of Topological Photonics is growing rapidly and is quite specialized, so the presentation here is necessarily incomplete – but sufficient to inform the reader for the main areas of current research. The presentation primarily concerns quantum systems ˙ however, some of these topological effects can also occur in classical optical systems [219].

3.2 Photonic crystals, waveguides, and coupled resonant cavities

3.2.1 Photonic crystals

In 2005 Haldane and Raghu proposed theoretically an analogue of the anomalous quantum Hall effect in optical systems [90, 200]. The system they studied was a photonic crystal, specifically a gyromagnetic one, where time-reversal symmetry is broken due to magneto-optical effects. Three years later, the idea was verified by Wang et al., who provided realistic material designs [262] and experimental observations [263]. These studies raised a plethora of subsequent investigations, theoretical [87, 64, 120, 160, 228] and experimental [135, 203, 88].

A photonic crystal [110] is an optical nanostructure in which the electric permittivity, hence the refractive index, change periodically in space. Photonic crystals affect

[3] i.e., symmetry under interchanging two distinct types of lattice sites.

photons the same way atomic lattices of solids (crystal structure) affect conductivity electrons. Photonic crystals can be constructed in 1D, 2D or 3D versions. 1D ones can be made of thin film layers deposited on each other. 2D ones can be made by photolithography, by drilling holes or arrange nanorods periodically in a suitable substrate. For 3D ones, fabrication methods include stacking multiple 2D layers on top of each other, drilling at different angles, instigating self-assembly of spheres in a matrix and dissolving the spheres, and techniques with lasers.

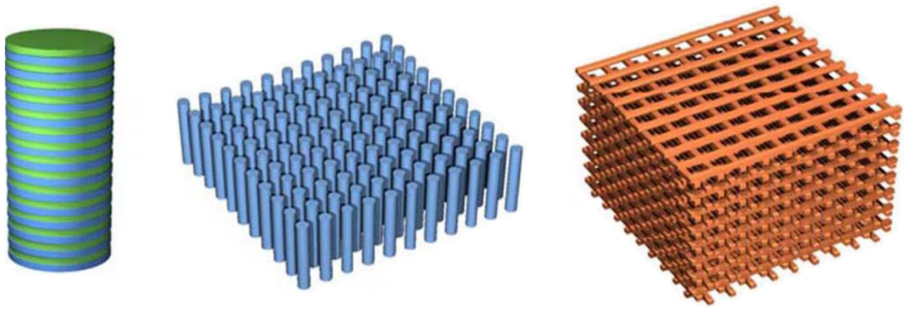


Figure 3.1: Schematics of representative lD, 2D and 3D photonic crystals.

The notifying feature of a photonic crystal is the periodicity of dielectric material (and thus the refractive index) along one or more axes. When an optical wave encounters a change in the refractive index of the media where it travels, always undergoes reflection. In a photonic crystal, the reflected wave interferences with the incident wave, constructively at some frequencies, and destructively at others. At the frequency ranges where the intereference is constructive, the wave propagates in the crystal ˙ these are allowed photonic bands. At the frequency ranges where the intereference is destructive, the propagation is imposible ˙ these are forbidden bands. The forbidden bands constitute the energy gaps required for the nontrivial topological states[4]. In conclusion, in photonic crystals, the periodic change of refractive index results eventually in photonic bands formation.

A photonic crystal consists of a periodic array of nanostructures (e.g., nanorods or holes) with high refractive index, embedded in a medium with lower one. The electromagnetic field tends to concentrate inside the high index structures ˙ therefore, the nanostructures in the ambient medium act as lattice sites where the field tends to localize. The hopping amplitudes of this localized field due to evanescent coupling between the sites, can be changed by changing the distances between the structures and/or their refractive index. In this way, usual solid-state structures and discrete hopping models such as an 1D SSH model or a 2D graphene-like honeycomb structure can be set with photonic crystals.

In the pionneer works [91, 200] it was predicted that could exist 2D photonic crystals with Hamiltonians of nonzero Chern numbers. These systems would have unidirectional edge states extremely stable and durable, even in high levels of impurities or disorder ˙ states for propagation in the other direction, permitting scattering of the

[4] This requirement has already been mentioned in p. 37,
and also when discussing the quantum Hall effect, in p. 54.

photons, would not exist. Bringing together two such photonic systems with different topological phases, then boundary states would be trapped along the interface, between bulks with different Chern numbers. These predictions were confirmed experimentally in subsequent studies [262, 263].

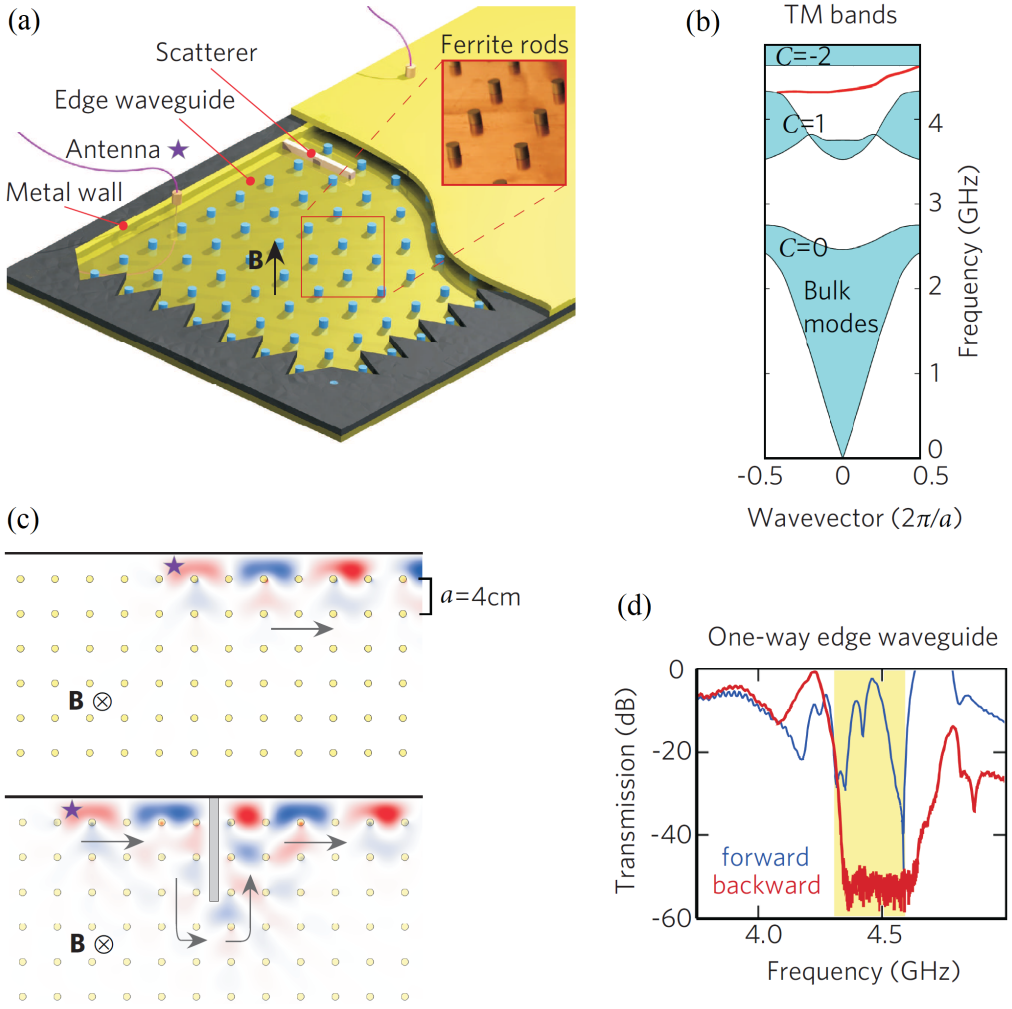


Figure 3.2: (reprinted from [161]). (a) Schematic of the experimental setup for observing the unidirectional edge state between the metal wall and the gyromagnetic photonic crystal. (b) The band structure of the system. The unidirectional gapless edge state between the second and third bands of non-zero Chern numbers is indicated. (c) Simulated field propagation of the unidirectional mode and its topological protection against large obstacles. (d) The measured transmission coefficient in the unidirectional edge of the waveguide.

Specifically in [263], the experiments used a 2D square lattice photonic crystal, consisting of an array of gyromagnetic ferrite rods confined vertically between two metallic plates, mimicking the 2D transverse magnetic (TM) modes, Fig. 3.2a. To avoid

radiation losses into the air, a metallic wall was set at the two sides of the crystal. In absence of external magnetic field, the band structure of the system has a quadratic point-degeneracy consisting of a pair of Dirac cones [47] that connect the second and third TM band. When a uniform static magnetic filed is applied[5], then TRS breaks, and in the magnetic permeability tensor appear anti-symmetric imaginary off-diagonal terms. The quadratic degeneracy is lifted up and a bandgap is formed between the second and third bands, both of them having non-zero Chern numbers. In this bandgap, at around 4.5 GHz, an edge state also appears, Fig. 3.2b, which is striclty unidirectional and has positive group velocities. Numerical simulations confirmed that a source inside this waveguide emits only forward in the bulk frequency gap, and there is no scattering even when the wave encounters quite large obstacles, Fig. 3.2c. The measurements of the trasmission coefficient, Fig. 3.2d, also shows that backward reflection is more than five orders lower than the forward propagation – and more imortant this low reflection still holds even after inserting large obstacles in the way of the wave, Fig. 3.2c. In fact, when the wave creeps on an new interface due to an obstacle, then new unidirectional edge states are formed, providing a path for light to circumvent the obstacle. Just this is the topological protection provided by a photonic crystal with non-zero Chern numbers !

It is noted that unidirectional waveguides of other types also exist, for example based on conventional non-reciprocal materials ˙ however, they lack the robustness of the topological protected structures like the above. Unfortunately, optical materials have very weak response in magnetic processes, making the realization of optical topological structures challenging. This is why topologically protected states were first observed in microwave frequencies : in this spectral range the magneto-optic effect required for breaking the TRS is strong – but now have been observed in other frequency ranges too. A topological insulator without the need for gyromagnetic effect was realized experimentally for first time in [203]. In that study an array of waveguides was used, coupled evanescently, and arranged in a graphene-like honeycomb pattern. The light propagates monotonically along the z-axis, thus the z coordinate is considered as an effective time variable. To break the effective TRS as required, the waveguides had helical shape.

Afterwards, similar topological phenomena observed and realized in a plethora of other systems [87, 88, 46, 63, 155, 135], mainly in optical cavities, quasicrystals, metamaterials, and coupled optical oscillators. Also, another class of systems exhibiting behavior of topological insulators are the optical resonators [88, 171], to be discussed next.

3.2.2 Optical resonators

Optical resonators [98] are resonant cavities, usually having the form of rings, in which optical fields can be stored for long times as configurations of resonant standing waves or circulating waves. Optical resonators can be coupled to each other or to waveguides by evanescent coupling. Bringing two such cavities close enough, then the evanescent field from one cavity penetrates a small distance into the other and allows tunneling of the field into the other cavity ˙ in this way, by evanescent coupling two optical resonators

[5] with a strength about 0.25 T.

can be coupled together or to waveguides. This inter-cavity coupling can be fabricated to achieve desired values with very high precision. Additionally, ring resonators have the capability to accept optical excitations that circulate either clockwise or anticlockwise, allowing to model various two-state systems such as electron spin.

Photons in an array of coupled resonators are similar to electrons in an array of atoms in solids. The photonic coupling between the resonators can be adjusted to form topologically non-trivial frequency gaps with robust edge states. Photonic analogues of the integer quantum Hall effect have been achieved using both static and time-harmonic couplings that simulate the electron's behavior in a uniform magnetic field. In these structures, the TRS breaking can be imposed by accurate time-harmonic modulations ˙ then, unidirectional edge states (waveguiding) immune to disorder can be realized at optical frequencies.

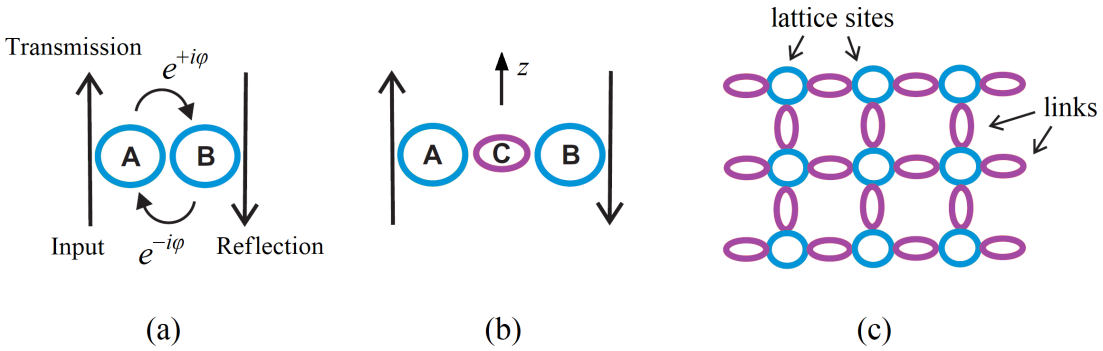


Figure 3.3: (adapted from [227]). Operation principle of optical ring resonators. (a) Two optical ring resonators coupled evanescently to each other and to external waveguides. (b) Same as (a) but here an auxilliary resonator is used to couple the original two in order to provide more control. (c) A 2D array of coupled resonators, based on pattern (b). The circular resonators act as lattice sites, while the elliptical ones serve as links that can be setup to provide hopping values to simulate different lattice models. For input and output, waveguides can be coupled to any of the lattice sites.

In Fig. 3.3 the operation principle of coupled optical resonators is sketched. In Fig. 3.3a it is shown a pair of resonators, A and B, coupled to each other and to external waveguides. Due to the evanescent coupling, a hopping phase ϕ occurs as the field oscillates from A to B and vice versa. In this structure, the phases in each direction are precisely equal in magnitude and opposite in sign. In Fig. 3.3b it is shown an improved version. An auxiliary resonator C has been added to provide more control and flexibility. Shifting C along z direction changes the linkage length between A and B ˙ this introduces an asymmetry between the hopping phases of the amplitudes for the left-to-right and right-to-left moving of the wave. In this way, the phase accumulated as the wave moves along a closed path[6], $e^{i\phi} = e^{-ie/\hbar c \oint \mathbf{A}\cdot d\mathbf{l}}$, can simulate an artificial gauge field $\mathbf{A}$, thus providing a means to break the TRS. Moreover, the tilting of A and B resonators against

[6] It is noted that this is the so called *Aharonov-Bohm phase* [161], not the Berry phase.

C gives the capability to impose or break the chiral symmetry[7]. With this flexibility, a variety of different possible topological systems can be fabricated.

In Fig. 3.3c it is shown a usual 2D structure of coupled resonators. The elliptical loops are the intermediate links, the circular ones are the main resonators. By adjusting the hopping amplitudes caused by the links, a plethora of solid-state lattice models can be simulated optically. An interesting case is to arrange the links in order the phases to vary from row to row. In a such configuration the phases on the upper and lower halves of a closed loop do not cancel · this provides an artificial gauge field with nonzero fluxes inside the areas enclosed by the loops.

It is emphasized that a photon does not interact with magnetic fields, but it also accumulates a phase change after sweeping a closed loop[8]. Therefore, the idea to impose time-harmonic modulations to indirectly create an artificial gauge field corresponding to the magnetic field, and then use it to break TRS, is crucial for optical resonators[9]. The pioneer work on this was done by Hafezi et al. [87, 88], who eventually succeed to show that certain robustness against particular types of disorder in optical resonators can still be achieved due to the topological features of the phase arrangements [171]. Fang et al. [64] proposed theoretically how to break TRS and eliminate backscattering using spatially-coherent time-domain modulations. Unfortunately, it is challenging to achieve accurate and coherent time-harmonic modulations of a large number of resonators in optical spectrum. Rechtsman et al. [203] converted the modulation from the time domain to the spatial domain, succeeding to demonstrate experimentally the photonic analogue of the quantum Hall effect in optical frequencies. A more detailed and comprehensible discussion of the above issues, with the corresponding references, can be found in [161].

3.2.3 Waveguides topologically protected

It is remined that the boundaries between material domains with different Chern numbers support surface or edge states, topologically protected and strictly confined very close to the boundary. "Protection" means that disorder and perturbations (crystal defects, impurities, abrupt bendings in geometry) do not break the states, and the states do not undergo dissipation into the bulk region of the materials. Moreover, these states are unidirectional, that is they allow moving of charge carriers strictly in one direction only. In the same topological phase, states with opposite directions generally do not exist, thus the existing states are extremely stable as there is no available channel for a particle to backscatter into it.

Having these properties, such states are very usefull in Photonics for transport of light without scattering and losses. High-coherence optical quantum states are generally fragile and easily perturbed by interactions with the environment. These perturbations

---

The Berry phase concerns quantum status vectors $|u\rangle$. The Aharonov-Bohm phase concerns the magnetic potential **A** and is proportional to the magnetic flux enclosed by the loop [227].

[7] i.e., symmetry under interchanging two distinct types of lattice sites.

[8] In fact the same phase as the Aharonov-Bohm phase of electrons moving in a uniform magnetic field.

[9] The idea to create effective magnetic fields for neutral particles [48] using artificial gauge fields was first studied in optical lattices [105]. Later on, similar gauge fields were also studied in Optomechanics [213] and radio-frequency circuits [107].

can decohere wavefunctions, that is the relative phases in quantum superpositions become random. As a result, quantum effects such as interference and entanglement are reduced or eliminated. Therefore, waveguides topologically protected are very useful to transport optical quantum states without decoherence. In Fig. 3.4 is shown a schematic of an 1D waveguide for optical transport, topologically protected. In the same manner, topologically protected states on 2D interfaces between 3D bulks can be used to restrict photons on a plane in order to reproduce physical phenomena in two dimensions.

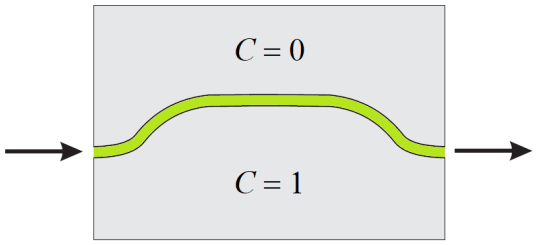


Figure 3.4: Operation principle of a topologically protected waveguide. Two materials with different Chern numbers (different topological phases) are in contact, inducing a state confined on the boundary. The boundary is used as a waveguide for optical transport, with the topological protection of the state to prevent degradation from environmental perturbations.

Many optical devices like waveguide splitters, signal switches, directional filters, have also been proposed or realized in topologically protected versions. Besides them, topologically protected surface states could be used in precision optical measures and sensing applications.

A basic concept in Quantum Optics, Quantum Information Processing, and other such disciplines is *entanglement*. Two- or multi-particle states that are entangled cannot be factored to a product of single-particle states. Entangled states contain correlations between the particles which are stronger than any classical correlation [83]. Thus, the design and realization of waveguides capable to topologically protect entangled states and quantum correlations is of great importance [27, 261, 172, 204].

3.2.4 Topological lasers

All the devices discussed so far are *passive*, meaning they operate without requiring external energy. In contrary, *active* photonic devices pump energy from an external source and they are used to amplify signals or produce nonlinear effects. Topological materials have applications in this case too [183] – and the main case of active photonic device with topological properties is the topological insulator laser [94, 11].

The basic concept of a topological insulator laser is to use an array of coupled ring oscillators and to pump the resonators only on the boundary of the array. The system

is topologically protected, thereby suppressing the loss of the signal in the bulk and preventing defects from stopping the energy propagation along the boundary.
This was tested in [94, 11] · the result was a high-coherence single-mode laser in which the ratio of laser intensity to pump intensity, and the output coupling efficiency, were significanlty higher than comparable nontopological lasers. Specifically, the output gave a peak near 1550 nm, in the wavelength region used for applications in standard telecommunications. The laser process takes place mainly at the boundary of the structure. Imperfections in the boundary do not influence the output of the topological laser · the energy flux penetrates into the bulk just to circumvent the defect, without scattering, and again returns to the boundary. This is not the case in a common (nontopological) laser, in which the defects cause scattering, thus disrupting the energy flow and the emission in their vicinity.

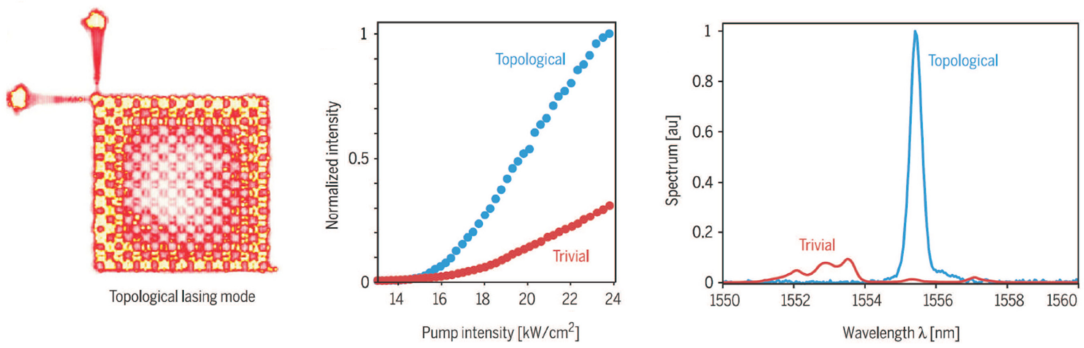


Figure 3.5: (reprinted from [11]). A topological insulator laser.
(a) Resonator array, pumped at the boundary. At the upper left lies the output port. The energy remains near the boundary, without leakage in the bulk, thus resulting to higher efficiency. (b) Output intensity of the single-mode topological laser and the maximum output mode of a comparable trivial laser. The first is greater by a factor five or more. (c) Output bandwidth of the single-mode topological laser and the multi-mode of the trivial laser. The first is significantly narrower.

A plethora of topological lasers have been proposed and demonstrated. Of the first successful attempts was a topological laser based on gyromagnetic photonic crystals [8]. Many other variations realized after that, like topological vertical cavity surface emitting lasers, topological quantum cascade lasers etc [278, 4, 230, 157]. Due to their stability, topological lasers can have arbitrary shapes, and imperfections in their construction have minimum influence in their performance. More speciallized versions are also in development · among them are nanocavity topological lasers [183, 93, 190] which exhibit high speed and low power consumption, and are suitable to be incorporated in integrated nanophotonic circuits.

3.3 Photonic walks

Quantum walk (QW) is a powerful technique for creating quantum algorithms, and for simulation of complex quantum systems. The quantum walk is the quantum version of the classical random walk, which is based on the "tossing of a coin" to determine the direction of the next step. According to Quantum Mechanics, the evolution of an isolated quantum system is deterministic : randomness occurs only when the system is measured and classical information is obtained[10]. The aforementioned coined model evolves at discrete time steps on a discrete space, represented by a graph. Two other main versions of QWs are also available : a coinless version known as *staggered model*, which uses an evolution operator defined by partitioning the vertex set, and a continuous-time version [193]. In this Section, only discrete QWs will be examined, in which a discrete step is taken at times $t = mT$, T being fixed.

Quantum walks of photons on discrete lattices [239] is another version, particularly interesting because it is quite simple in experimental realization and can simulate a plethora of phenomena.

In a classical random walk [128], a particle lies on a discrete lattice, which here it is taken to be 1D for simplicity, and at each multiple of a discrete time T the particle moves from its current position to one of the adjacent sites, left or right. It can be proved [205] that the probability of being at lattice site n after m time steps is given by a binomial distribution. When m is large enough, then, according to the central limiting theorem this probability becomes approximately gaussian. The width of the distribution, measured by the standard deviation, grows proportionally to the square root of time, $\sigma = O(\sqrt{m})$. Such a time dependence occurs in diffusion processes, and it is called *diffusive spread*.

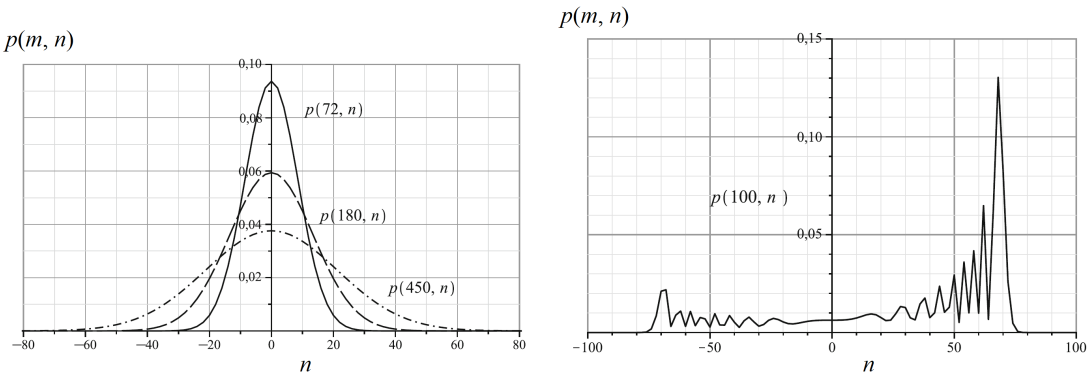


Figure 3.6: (a) Probability of a classical random walk. The standard deviation is diffusive, $\sigma = O(\sqrt{m})$. (b) Probability of a quantum walk. The standard deviation is ballistic, $\sigma = O(m)$. For the "coin" operator and the initial conditions used in this case see [193].
n is the lattice site, m the time step.

[10] This is the reason the name "quantum random walk" is rarely used.

Quantum walks exhibit a very different behavior [1, 118, 255, 259]. In a quantum walk, instead of a classical particle there is a wavefunction. Speaking of the 1D case, at each time step the wavefunction spreads in both directions simultaneously (left and right). After a few steps, an ensemble of paths has been formed that can be swept from the initial site to other nearby lattice sites˙ this causes quantum interference between the different possible paths. The wavefunction amplitudes for moving left or right can be determined by a random variable named the *coin variable*. In fact, coin variable is a two-component vector, with its components concerning the amplitude for leftward and rightward steps. At each step, the amplitudes for the left and right steps are determined by a random process ("coin tossing")[11] and vary. To describe the process, the Feynman path integral method [210] can be used, which gives quantum amplitudes by summing over ensembles of classical trajectories. Whatever the case is, the amplitudes $a(n)$ for the particle to be at each site n evolve deterministically over time, and interfere with each other in a complicated way˙ thus, the final probability distribution is much more complex than in the classical random walk. The probability a site n at time m to be occupied is given as $p(m, n) = |a(m, n)|^2$. An indicative example is shown in Fig. 3.6. It is evident that the distribution is not gaussian˙ indeed it tends to be smaller in the middle and larger at the ends[12]. Near the ends the probability has a peak that is more than 10 times larger than the values at the origin. This suggests that the quantum walk has a *ballistic behavior*, meaning that the particle can be found away from the origin as if it is in a uniform rightward motion[13]. In this case, the standard deviation grows proportionally to the time, $\sigma = O(m)$. The asymmetry in the distribution is due to the choice of coin variable used˙ with other "coins", or different initial conditions, more symmetric walks can be produced.

Due to their ballistic spreading, the quantum walks spread faster than any classical random walk, and can be used to physically realize fast quantum search algorithms. For this reason, QWs are of great interest in Quantum Information Processing and Quantum Computing. An important characteristic in many quantum computation algorithms is that the superposition principle, interference, and other quantum properties result to the so called *quantum speed-up* of algorithms [175, 170] – and quantum walk processes can be used to model universal quantum computers [5, 193, 254]. Besides these, it is noted that Berry phase and holonomy have important role in quantum walks [197].

A basic theoretic ingredient in discrete QWs is the time evolution operator $\widehat{U}$, which pushes states one time step forward, $|\psi(t)\rangle \rightarrow |\psi(t+T)\rangle = \widehat{U}\,|\psi(t)\rangle$. If $\widehat{U}$ is known, an effective Hamiltonian can be defined for the walk[14], if the type of the system permits it. For a closed, conservative system the time evolution operator is known to be [177]

$$\widehat{U}(t) = \exp(-i\widehat{H}t/\hbar)\,. \tag{3.1}$$

[11] In practice, this is done with an appropriate operator, for example the so called *Hadamard coin* [193].

[12] Note that after an even number of steps only even positions can be reached (and similarly for odd positions after odd numbers of steps). The points with zero probability (n odd) are excluded and not shown in the graph.

[13] This reminds a freely propagating classical particle shot from a gun, hence the name.

[14] It is noted that in general the process as it is studied in Quantum Mechanics works in the opposite : the Hamiltonian of the system is known and the effort is to obtain the time operator from it [177].

Then, an effective Hamitonian for the QW is defined as[15]

$$\widehat{H} = \frac{i\hbar}{T} ln\widehat{U} \,. \tag{3.2}$$

In this case, $\widehat{U}$ and $\widehat{H}$ define in fact a Floquet system[16], in which the driving period is the discrete time step T.

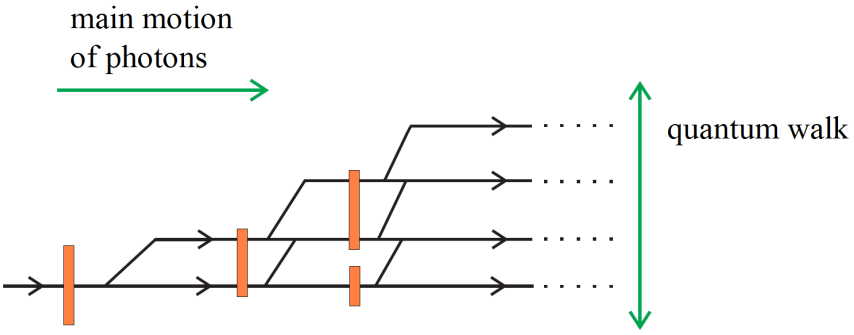


Figure 3.7: Operation principle of an optical quantum walk.
The "coin tossing" operation is realized by rotating polarization states using half-wave plates (orange rectangles), then separating polarization components using birefringent elements (at the splitting points of the lines). The rotation sizes alternate to create nontrivial topological phases. The photons move primarily rightwards, and perfom a quantum walk in the vertical direction.

It has been shown experimentally [36, 127] that the behavior of Hamiltonians with nonzero Chern number can be simulated by 1D quantum walks of photons in an optical system ˙ and that bringing together two such systems with different Chern number, topologically protected optical boundary states can be formed, as expected from the theory of the topological insulators. This was achieved using 2D arrays of beam splitters and phase plates, with the photons moving along one axis and performing an 1D quantum walk along the perpendicular axis, Fig. 3.7. In fact, this is a Floquet system – and indeed one in which spatial variations in the z direction are used to drive periodically the system. In this case, the spatial variation is realized using a so called split-step walk, in which each step of the walk consists of two separate substeps with different translation and "coin tossing" processes [36, 127]. Another technique to accomplish similar effects in an 1D optical structure has also been proposed in [226].

[15] It is remined that the exponentials and logarithms of operators are defined through their power series expansions.

[16] Floquet systems are those in which the Hamiltonian is periodically driven, for example by varying periodically in time the parameters of the system. Often, if the particles being described are moving monotonically in some direction (for example the z-axis), then this direction can be used as a substitute for time, and the driving can be accomplished by a periodic spatial variation of the system. This trick is sometimes used in photonic crystals and optical resonators for breaking the TRS.

4. Extraordinary Optical Trasmission through subwavelength apertures

4.1 Introduction

Optical elements and structures are generally bigger than electronic ones. For example, nowadays fabricating transistors smaller than 20 nm is common place, but to construct optical devices smaller than 50 nm is very difficult. The main factor preventing the shrinkage is the diffraction limit of light, which roughly states that light cannot be guided or stored in structures smaller than half of its wavelength. Although this is a fundamental principle for the wave phenomena, over the past twenty years a number of ways have been found to circumvent it – a prominent one of them being the main subject of this text[1]. A milestone to this problem was the discovery of *Extraordinary Optical Trasmission (EOT)* by Ebbesen et al. in 1998 [55]. They demonstrated that light can pass through holes with dimensions much smaller than the operating wavelength of light using excitations of plasmonic waves in the nanometer range. This opened up a road for new applications and construction of optical devices much smaller than the operating wavelength of light.

Due to its significance for applications, the passage of light through subwavelength holes in an opaque screen has been a topic of intense research for over a century. There is a wealth of important works studying single holes or arrays of holes, the earliest of which date back to Rayleigh's interpretation of diffraction in metal gratings [202]. The unusual EOT phenomena were observed for first time experimentally in hole arrays on metallic films, where the holes were significantly smaller than the operating wavelengths [55]. These first studies gave impetus to an extensive investigation of the transmission properties in a wide variety of cases, namely various shapes of holes, arrays of holes in various formations, holes surrounded by periodic structures etc. All these concern mainly nanostructures, as the mechanisms of EOT are in the nanoscale and practical requested is the construction of optical devices in nanoscale.

Despite the extensive research, the exact mechanism of EOT is not understood in all its details, and is still a topic of debate [77, 49]. In this chapter the basics of EOT will be discussed, using the most generally accepted interpretations. The enhancement of transmission in a single hole, and also in hole arrays is examined, and then how directional control of the transmitted light can be achieved using corrugations around the holes. A variety of cases are presented, and the role of SPPs and LSPs in these pheno-

[1] It will be examined exhaustively in a later chapter.

mena is discussed. All these give a background to better understand and appreciate the innovative method to overcome the subwalength transmission limit, presented in a later chapter.

4.2 Elements of Diffraction by subwavelength apertures

Because of its wave nature, when light passes through an aperture induces diffraction. Even for the simplest geometries this is a complex phenomenon, and can be described by various approximations provided by the classical Diffraction Theory [106, 33]. A very common case, studied extensively due to its theoretical tractability, is a circular aperture in a thin metallic film, Fig. 4.1. For an aperture with radius much larger than the operating wavelength, $r \gg \lambda$, the problem is treated by the scalar diffraction theory of Kirchhoff [106]. This theory is based on the scalar wave equation, so the polarization of light does not taken into account. In the far field, the transmited intensity[2] of a plane wave impinging normally on the film, is given by the relation[3] [80]

$$I(\theta) = I_0 \frac{(kr)^2}{\pi} \left| \frac{J_1(kr \sin\theta)}{kr \sin\theta} \right|^2 . \tag{4.1}$$

In (4.1), I_0 is the total incident intensity illuminating exactly the hole area πr^2, $k = 2\pi/\lambda$ is the wavenumber of the incident wave, θ is the angle between the hole axis and the direction of exiting radiation, and J_1 is the Bessel function of first kind.

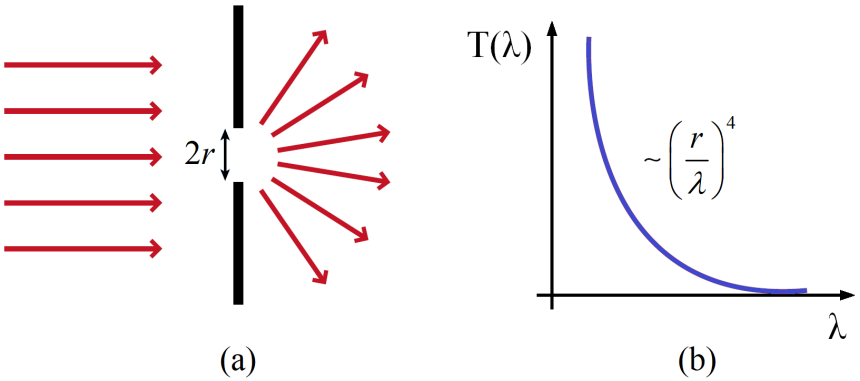


Figure 4.1: (a) Diffraction of light through circular hole in an infinitely thin metal film. (b) Typical transmission coefficient (Bethe's theory).

[2] The intensity I can be defined to be either a radiometric or a photometric quantity. As a radiometric quantity, I is the *radiant intensity* : it concerns radiation per unit solid angle in a given direction, measured in W/sr. As a photometric quantity, I it is the *luminus intensity* : it concerns visible light flow from a light source per unit solid angle in a given direction, measured in cd (candelas). It is 1 cd = 1 lm/sr, and 1 W = 683 lm (lumen) of monochrome light at 555 nm.

[3] This is known as the limit of Fraunhofer diffraction.

Eq. (4.1) describes the well known Airy pattern[4], appearing when diffraction takes place from such circular holes : a central bright spot, surrounded by concentric rings, of decreasing intensity away from the center. The alternating of bright rings with dark ones is due to the constructive and destructive interference of light exiting from the aperture. Similar patterns appear in apertures with other schemes (rectangular, single slit, double slit) [80]. Among others, an important quantity is the transmission coefficient, defined as the ratio of total transmitted intensity to I_0,

$$T = \frac{\int I(\theta)\, d\Omega}{I_0} \,. \tag{4.2}$$

In the examined case it is[5]

$$T = \frac{1}{\pi} \left| \frac{J_1(kr \sin\theta)}{\sin\theta} \right|^2 \propto \left| \frac{r}{\lambda} \right|^2 \,. \tag{4.3}$$

For apertures with radius much larger than the operating wavelength, $r \gg \lambda$, more accurate calculations result in relations like (4.3), in which the transmission coefficient tends to unity, $T \approx 1$.

In the other extreme case, namely subwavelength apertures with $r \ll \lambda$, a correct treatment, even rough, requires the Maxwell's equations. The reason is the following. Kirchhoff's theory is based on the condition that the electromagnetic field in the aperture is the same as if the opaque film was not present; this does not satisfies the boundary condition the tangential field to be zero on the (conducting) film. For large apertures, this crucial violation is acceptable because the diffracted field is relatively small compared to the field transmitted directly. But for subwavelength apertures, this approximation is insufficient, even in a rough treatment of the problem.

For the transmission of light through a subwalength circular hole, an exact analytical solution was derived by Bethe and Bouwkamp [22, 31, 32, 52]. In Bethe's theory, it is assumed that the film is perfect conductor (PEC) and infinitely thin, and the light intensity I_0 over the hole is constant. For normal incidence, the hole behaves like a small magnetic dipole and the transmission coefficient is found to be

$$T = \frac{64}{27\pi^2}(kr)^4 \propto \left(\frac{r}{\lambda}\right)^4 \,. \tag{4.4}$$

This equation holds for both TE and TM polarization, but for normal incidence of the wave. If the wave impinges on the hole at an angle, to describe correctly the transmission an electric dipole is additionally required [22].

---

[4] Not to be confused with the Airy function.

[5] It is reminded that

$$J_1(x) = \frac{x}{2} - \frac{x^3}{2^2 4} + \frac{x^5}{2^2 4^2 6^2} - \frac{x^7}{2^2 4^2 6^2 8} \cdots$$

Eq. (4.3) given by Kirchhoff's theory holds for $r \gg \lambda$, and the dependence of T with (r/λ) is quadratic. In comparison, (4.4) by Bethe's theory holds for $r \ll \lambda$, and the attenuation of T with (r/λ) is biquadratic[6]. The transmission through the subwavelength hole is very weak, as expected. As the wavelength of the impinging light becomes larger than the hole radius, the transmission through the hole is decreased very rapidly, Fig. 4.1b. Thus, the description of the phenomenon by (4.4) is intuitive and reasonable. However, Bethe's theory is based on two crucial approximations, already mentioned above : the metallic film is perfect conductor (PEC), and is infinitely thin. But a real metallic film does not fulfill these two requirements. The consequences are dramatic; the real transmission behavior is very different from that predicted by (4.4) – and in any case does not diminish monotonically with decreasing the kr.

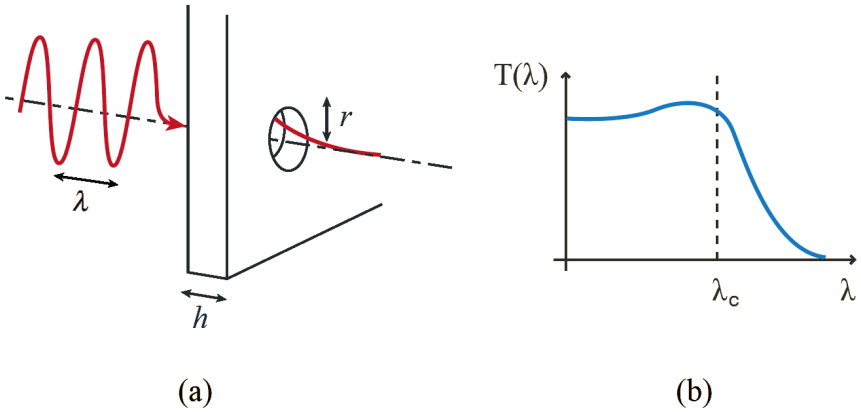


Figure 4.2: (a) Passing of light through a subwavelength hole on a real metallic film (thickness and conductivity are finite). The hole is characterized by a radius and a thickness, resembling a cylindrical waveguide.
(b) The transmission has a cutoff wavelength that defines two different propagating regimes.

A real metallic film has finite conductivity and thickness, and an aperture on it has a definite radius and depth, Fig. 4.2a. Hence the aperture behaves as a waveguide : the electromagnetic wave is confined within the aperture space, and the dispersion relation is defined by the radius and the depth. Also, as in waveguides, the lateral dimensions of the aperture define a cutoff wavelength λ_c . When the operating wavelength is smaller than λ_c, the impinging wave propagates through the hole; when it is larger, the wave attenuates exponentially, Fig. 4.2b. In real films, increasing the operating wavelength there is a gradual transition from the propagating to the evanescent regime, thus the cutoff wavelength does not have a sharply defined value [191].

[6] Also, the dependence $T \propto \lambda^{-4}$ agrees with the Rayleigh's theory of scattering by small objects [29].

4.3 Rudiments of EOT phenomena

The transmission of light through a subwavelength hole that on principle is very weak or not allowed at all, can be greatly improved by forming the screen with appropriate periodic structures (gratings or hole arrays). Keystone of this enhancement is the excitation of SPPs[7] and LSPs on the screen, and their coupling with these formations, leading to enhancing the light field near the hole. The SPPs perform tunneling through the hole and convey the energy of the field to the other side of the hole, where it propagates to the far field.

The coupling of SPPs with the periodic grating on the screen is imposed by a phase-matching condition, a well known process for exciting SPPs. The resulting transmission coefficient $T(\lambda)$ has peaks at the wavelengths where the SPPs are excited. At these wavelengths it is possible to be $T > 1$, meaning that more light can pass through the hole than the incident on the hole area; the reason is that light impinging on the metal screen near the hole is channeled through the hole via SPPs. This is an extraordinary transmission phenomenon, fisrt observed by Ebbesen et al. for a square lattice of circular holes on a thin silver film [55].

As a first example, in Fig. 4.3a is shown the transmission spectrum of a structure having EOT features [151]. The structure is a square lattice of circular holes, with diameter 200 nm for each hole and period 600 nm for the lattice, perforated on a gold film with thickness 100 nm. As can be seen, the transmission coefficient has peaks, indicating that more light passes from a hole in the presence of all the holes, than what is expected from a single hole on the film. The transmission spectrum has some remarkable features, discussed below, that are a manifestation of EOT.

For the propagating modes, a cutoff wavelength at ~ 340 nm is expected. But in contrary, above the expected cutoff value the transmission becomes stronger ! The trasmission coefficient has peaks at wavelengths longer than the cutoff value – and the highest peak is at 745 nm, more than twice the expected cutoff value. Also, each peak has a minimum value nearby, deeper at higher wavelengths. The transmission coefficient to have peaks that become stronger as the wavelength increases is a noteworthy characteristic of EOT.

Another important feature shown in Fig. 4.3a is the significant transmission efficiency. The values here are normalized to the intensity of the impinging wave on the surface and the units are arbitrary (a.u.), but in fact they are much greater than the values predicted by Eq. (4.4) of Bethe's theory [196].

A third feature of the EOT is the enhancement of the near-field on the irradiated surface. In Fig. 4.3b is presented a plot of the electric field intensity for the examined structure. As shown, the holes capture incident energy beyond their area and confine it in a small volume near them and near to the surface; this facilitates and increases the tunneling through the holes. At the entrance and exit edges of the holes significant enhancement of the local field takes place.

---

[7] For a reminder of what is the SPP see F/note 4, p. 115.

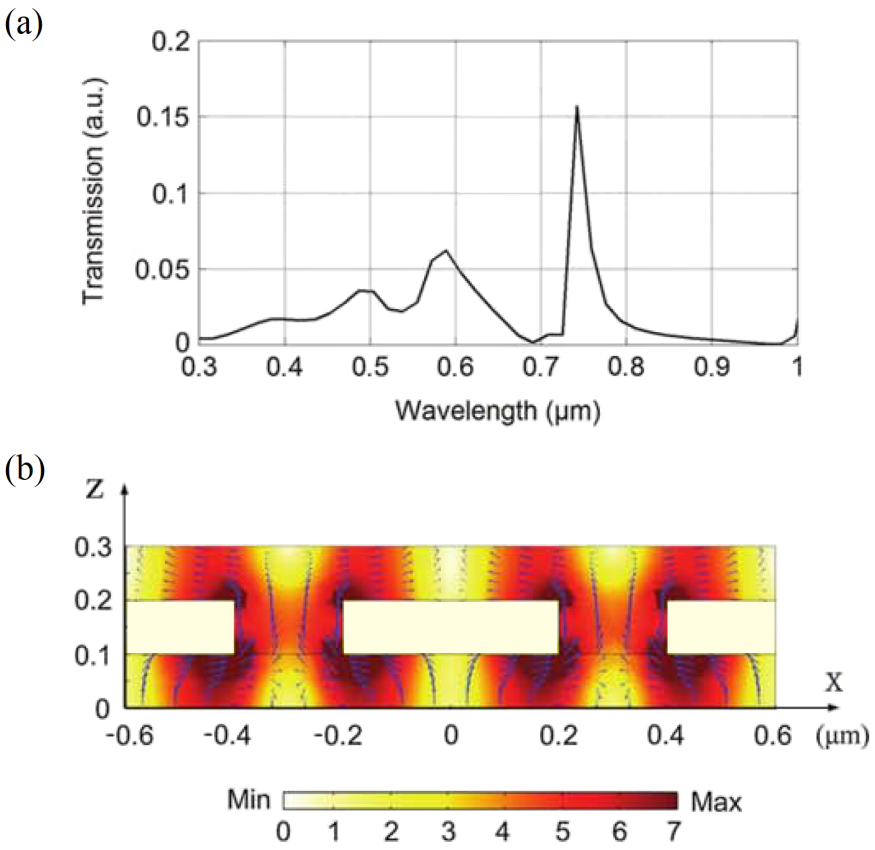


Figure 4.3: EOT through an array of circular holes on a gold film [151].
(a) Transmission spectrum for a normally incident plane wave.
(b) Electric near-field distribution and Poynting vector calculated at a resonance wavelength of 742 nm.

Summarizing, the transmission of light in the aforementioned structure exhibits the following noteworthy characteristics :

- longer resonance wavelengths than the cutoff wavelength,
- enhanced transmission efficiency,
- near-field enhancement in a very small volume.

The phenomena associated with these characteristics constitute the EOT. Although the EOT is not understood in all its details, many interpretations have been given based mainly on the excitation and propagation of SPPs on the irradiated surface [77, 49].

The crucial role of SPPs in EOT is revealed from the dependence of the peak positions of T on both the incidence angle of light and the lattice period. It is well known that coupling incindent light on a grating surface and exciting SPPs obeys the following phase-matching condition [163] :

$$\mathbf{k}_{SPP} = \mathbf{k}_{in} \sin\theta \pm n\frac{2\pi}{a_x}\hat{\mathbf{x}} \pm m\frac{2\pi}{a_y}\hat{\mathbf{y}}\,. \tag{4.5}$$

In (4.5), $\mathbf{k}_{SPP}$ is the wavevector of the excited SPP, $\mathbf{k}_{in}$ that of incident light, a_x is the grating lattice constant and $\hat{\mathbf{x}}$ the unit vector along x-direction, and likewise for the a_y and $\hat{\mathbf{y}}$. Obviously, θ is the incidence angle and n, m are integers denoting multiples of the grating period $G_i = 2\pi/a_i$, $i = \{x, y\}$ along $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$. Often the lattice is square, then it is $a_x = a_y \equiv a_0$ and $G_x = G_y$.

For phase-matching in a square grating lattice and normally incident light ($\theta = 0$), the transmission maxima occur at wavelengths given by the relation [79]

$$\lambda_{SPP}(n, m) = \frac{a_0}{\sqrt{n^2 + m^2}} \sqrt{\frac{\varepsilon_m \, \varepsilon_d}{\varepsilon_m + \varepsilon_d}} \, . \tag{4.6}$$

Eq. (4.6) also holds for a lattice array of square holes.
In the above, ε_m and ε_d are the relative permittivities of the metallic and dielectric media respectively, and a_0 is the lattice constant of the grating or the hole array. Eq. (4.6) concerns a metal-dielectric interface, where SPPs can be excited and propagate, a structure well studied in Plasmonics. ε_m and ε_d are generally complex, thus λ_{SPP} results complex too. The real part of λ_{SPP} concerns the resonance wavelengths, whereas the negative imaginary part concerns the non-radiative damping (due to absorption of the SPP into the metal). In practice, the experimental results differ somewhat from the predicted by (4.6). The deviations are due to scattering losses, differences between the property values in reality and those adopted in simulations etc. However, (4.6) is a good first approximation.

Furthermore, there is another effect that contributes to EOT. It is well known from Plasmonics that light at frequencies ω greater than the plasmonic frequency ω_p of a metal, can penetrate the metal. Ideally, if the permitivity is purely real, the metal is transparent to the incident light. In reality, for $\omega > \omega_p$ the imaginary part is small and the real part of the permitivity dominates strongly. Then the electric field can penetrate the metal at least to the order of the skin depth[8]. In these cases, the field penetration results in the hole to appear larger than its real size; the cutoff wavelength also appears longer and can be estimated by [82]

$$\lambda_c = \frac{\pi l \sqrt{\varepsilon_d}}{\arctan \sqrt{|\varepsilon_m / \varepsilon_d|}}, \tag{4.7}$$

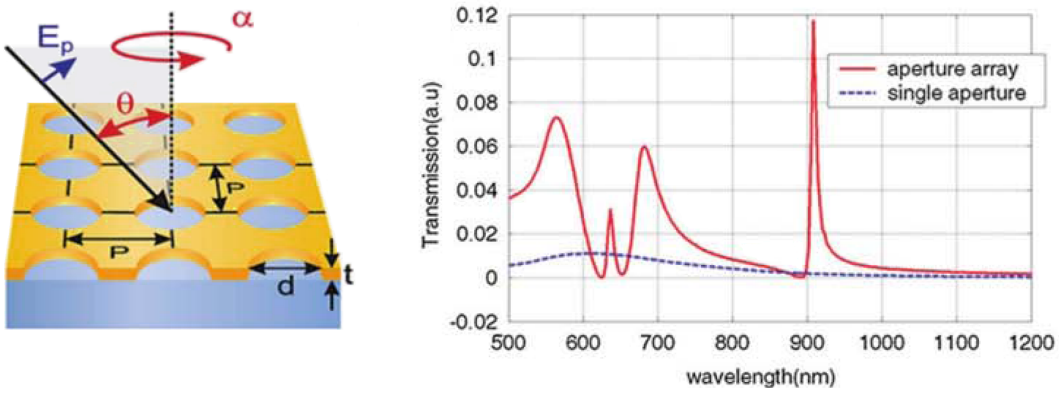
where l is the length of the hole or the thickness of the metallic film.
Eq. (4.7) means that the cutoff wavelength becomes longer (redshifts) due to the penetration of the electric field into the metal surface. For example, for a hole of diameter 270 nm on a silver film, irradiated by light at 750 nm, it is reported an increase $\sim$14% . The cutoff wavelength will increase even more, reaching up to $\sim 40\%$, if the coupling of SPPs is taken into account [82]. For the correct normalization of the transmission coefficient, the increase in the effective diameter of the hole must be taken into account, especially when studying holes with a diameter just below the cutoff diameter for a PEC screen.

[8] As it is pointed out in §4.6, in fact this happens due to local surface plasmons (LSPs) excited at the rim of the hole.

4.4 Transmission through hole arrays

The transmission spectrum of a hole array differs significantly from that of an isolated single hole. The transmission exhibits resonances only at specific discrete wavelengths [77]. In Fig. 4.4 it is shown a typical example : the transmission spectrum of a square array of circular holes vs that of a single hole, perforated on a gold film [151]. As can be seen, in the hole array, the transmission has much higher maximas, and there are sharp peaks and drops, indicating resonances. In contrary, in the isolated hole, the transmission is quite lower, it is smooth, and does not fluctuates. Evidently, the spectrum in the two cases is very different.

The transmission spectrum in a hole array is in general complex. The interpretation of the features exhibited involves the contribution of both the enhancing and suppressing effects [211]. An analytical model given indicates that the basic mechanism is the tunneling mediated by SPPs [166].



(a) (b)

Figure 4.4: Transmission through an array of circular holes [151].
(a) Geometry and coordinate system used.
(b) Transmission spectrum of an isolated single hole, and of a square array of holes. The incident wave is parallel polarized. The geometric characteristics are : $d = 200$ nm, $P = 600$ nm, $t = 100$ nm.

4.4.1 Influence of the number of holes, diameter and film thickness

In a hole array, the collective response is crucial factor for the EOT. As the number of holes increases the transmittance increases, until a saturation value is reached [211, 196]. The number of holes required to reach this value depends on the diameter of the holes : the larger the diameter, the faster the saturation is reached [196]. The maximum saturation value is related to the propagation length of SPPs, which is determined by the size of the holes [211].

The film thickness also influences the transmittance. It has been found that for quite thin films (a few tens nm or less) the transmittance varies differently than for a corresponding film relatively thick. In Fig. 4.5 it is shown an indicative study [151]. The incidence is normal, and three different cases are examined for the holes diameter (200, 275 and 300 nm). As it is seen, increasing the diameter the peaks are getting higher and they are shifted slightly to longer wavelengths, in overal enhancing the transmittance. As mentioned above, the complex behavior exhibited is due to a combination of enhancing and suppressing effects caused by the interaction and the tunneling of SPPs.

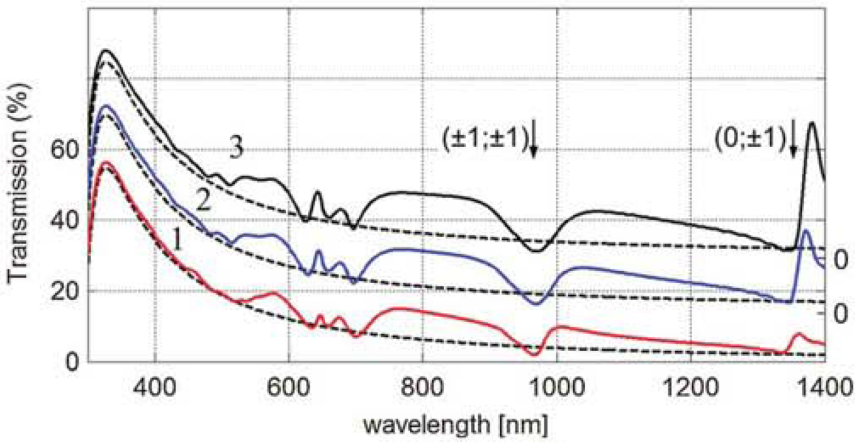


Figure 4.5: Transmission spectrum of a hole array [151].
Film thickness is $t = 20$ nm and lattice periodicity $P = 200$ nm.
Three different diameters for the holes are examined: $d = 200$ nm, $d = 275$ nm, $d = 300$ nm. The curves 2 and 3 have been shifted in the transmittance axis for clarity. The dotted curves correspond to a single isolated hole and are plotted for comparison.

The ratio of film thickness to the holes diameter also influences the transmittance [77]. In Fig. 4.6 it is shown the transmission spectrum of a hole array on a PEC, for a variety of ratios film thickness to holes diameter [151]. The results were obtained using the FDTD. As can be seen, for ratios less than 0.5 the curves have two maxima and in asymmetrical positions, and the transmission reaches a 100%. In this case the thickness is close to the skin depth (a few times larger), hence the metallic film is optically opaque. For ratios greater than 0.5 the curves have a maximum which still tends to 100%. For extreme ratios the transmission maximum still exists but it is quite low. In general, decreasing gradually the holes diameter while keeping constant the film thickness, there is a gradual transition from the two distinct peaks to a single peak; reducing the diameter of holes even more, this single maximum is attenuated. This behavior is reasonable as the smaller and sparser are the holes, the more difficult is for the light to pass through them.

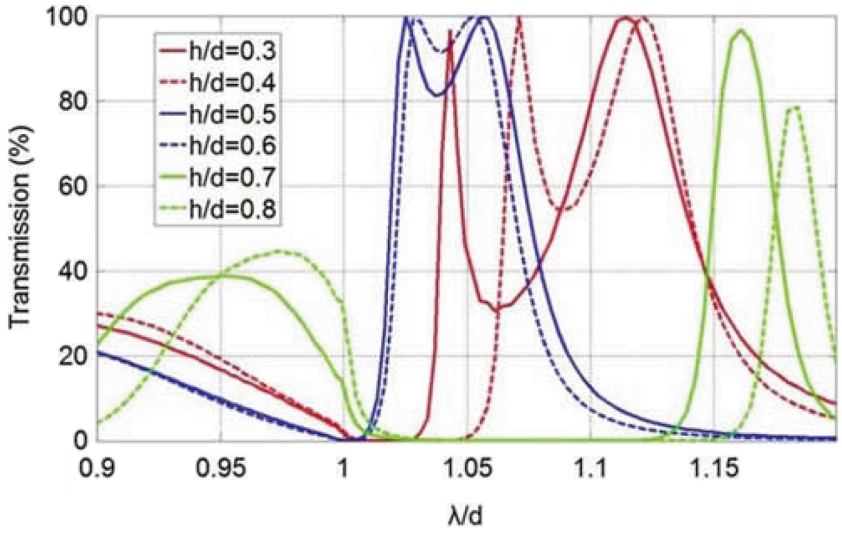


Figure 4.6: Transmission spectrum of a hole array on a PEC, for a variety of ratios film thickness to holes diameter.
The thickness h varies, the holes diameter d remains constant.
Diameter of holes is $d = 0.4$ μm and lattice periodicity $P = 1$ μm.
The wavelength interval is $900 \div 1200$ nm.

4.4.2 Influence of the polarization, the shape and size of the holes

It is known from Plasmonics that the size and shape of metallic particles influence strongly their optical response. In a similar manner, the geometric shape of a hole affects the characteristics of the transmission [131]. In arrays of rectangular holes, the transmission is also affected by the polarization of the incident wave. The influence the shape and size of the holes, and also the polarization of the wave, have on the transmission through hole arrays has been studied in a wide spectral range [112, 253]. Some results are presented below.

In Fig. 4.7 it is shown the transmission through a hole array for a variety of rectangular holes, and two cases of polarization; specifically, perpendicular (y-polarization) and parallel (x-polarization) to the long axis of the holes [253]. Evidently, the polarization of the incident wave affects the transmission significantly (note the different scale in the two graphs).

Fig. 4.7a concerns the case for an incident wave y-polarized. Increasing the aspect ratio of the holes, the peak located initially at ~ 720 nm grows and broadens significantly, and subject to redshift. Another peak, initially at ~ 600 nm, also grows and redshifts but not so much.

In Fig. 4.7b the case of x-polarized light is examined. Now as the aspect ratio of the holes increases, the peak of the resonance wavelength decreases and subject to blueshift. Compared to Fig. 4.7a, the scale of the y axis is much smaller, indicating that the aspect ratio of the holes plays a crucial role in the transmission magnitude and clearly induces polarization anisotropy. In general, in a hole array with elongated holes,

the transmission behavior is determined by both the localized effects due to the single hole and the collective effects caused by the periodic holes arrangement.

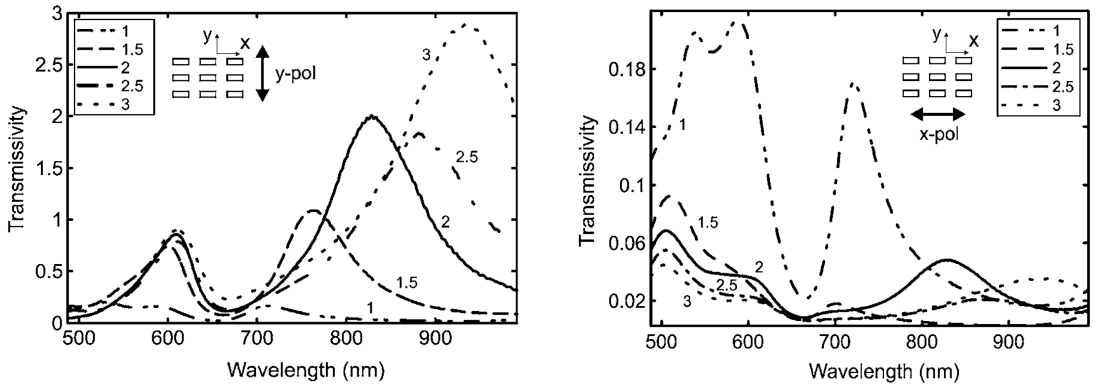


Figure 4.7: Transmission in hole arrays for two cases of polarization and misc aspect ratios of the holes dimensions (hole area remains constant) [253]. The array is contructed on a gold film of thickness 200 nm and periodicity 425 nm. (a) y-polarization : increasing the aspect ratio the transmittance increases and the light redshifts. (b) x-polarization : increasing the aspect ratio the transmittance decreases and the light blueshifts.

Fig. 4.8 concerns the influence that has to the transmission the shape and area of holes of an array [131]. Three types of hole arrays are examined : one with circular holes, and two with rectangular ones, of different area. The circular holes have a diameter of 190 nm. In all cases the holes are arranged in a square lattice of periodicity 425 nm. The results are normalized by the area occupied by the array holes.

In all the three considered cases the holes exhibit EOT behavior. The rectangular holes give peaks much more higher than the circular ones, even in the case of the rectangular holes 75×225 nm^2 which have quite smaller area than the circular holes. Evidently, the diameter of the holes determines both the transmission magnitude and the resonance wavelength. Observe also that as the rectangle deviates more from a square (here the third case), the maxima and the redshift of the peaks become greater.

In Fig. 4.9 it is shown the energy density inside a hole of a hole array [151]. The details are not of much importance; informationally it is reported that the geometry parameters of the array are : thickness 100 nm, period 300 nm, and slit width 50 nm. The imporant observation is that the graph indicates resonance for the SPPs on both sides of the array. In such a case, pairs of intensity spots are concentrated around the holes, resulting to high and localized intensity gradient on the external surface, mainly near the holes, thus enhancing locally the field. A similar case of near-field enhancement was also presented in Fig. 4.3b. If both sides of the holes in the metal film are surrounded symmetrically by an appropriate medium and have the same dielectric constant, then it is easy to achieve resonance for the transmission between the two sides because the SPPs on the two interfaces coincide due to symmetry [136].

As seen in Fig. 4.9, the intensity of the field is confined very close to the interface between the metal-dielectric, and decreases very rapidly with the distance from the interface. Inside the hole the field is very high, indicating that the incident energy is tunneled through the hole. Other studies indicate that to the EOT of hole arrays, besides the SPPs, contribute also the evanescent waves and normal guided modes [256]. The exact way all these factors contribute to the EOT is still under investigation.

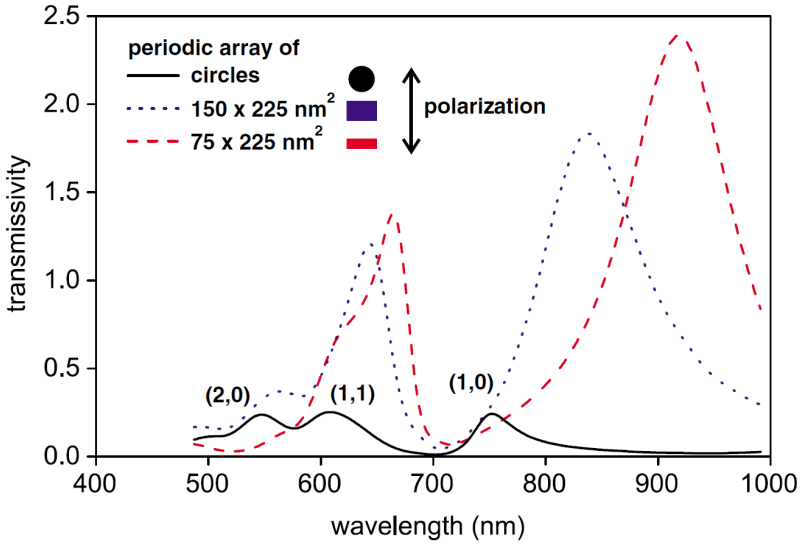


Figure 4.8: Transmission spectrum for hole arrays with different hole shape and area [131]. In each case the holes are perforated on a gold film with thickness 200 nm deposited on glass, in a square lattice with periodicity 425 nm.

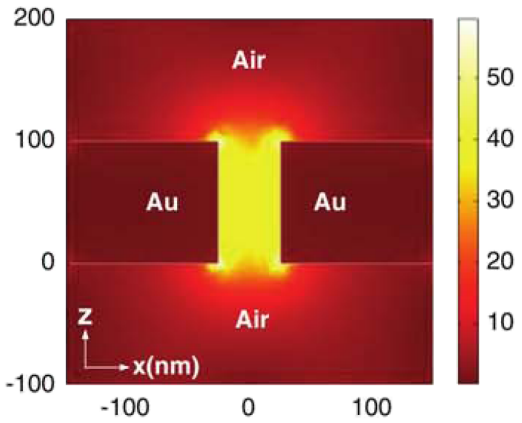
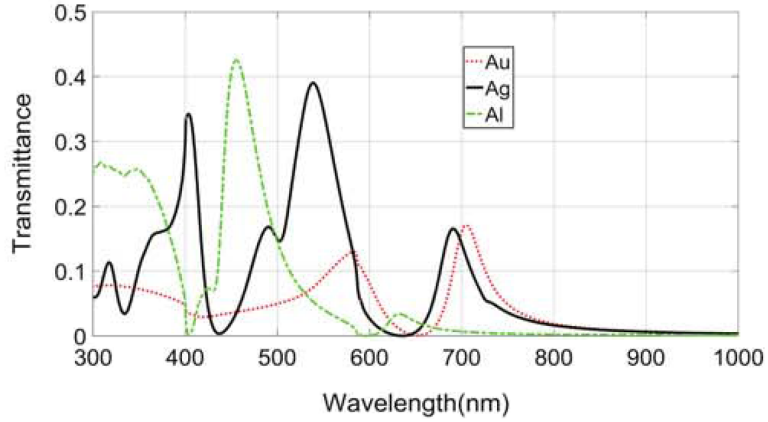


Figure 4.9: Normalized energy density inside a hole of a hole array (the scale is logarithmic).

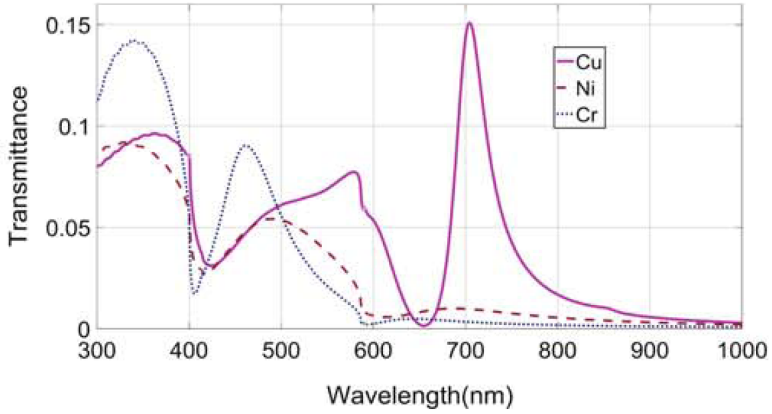
4.4.3 Influence of the type of metal

It is well known from Plasmonics that the properties of metals[9] determine their optical response and the behavior of SPPs on them. As SPPs is the basic mechanism of EOT, it is expected the type of metal used for the film to affect the transmission. Indeed, in the optical regime, only noble metals such as Au, Ag and Cu enhance significantly the transmission; metals like W that cannot sustain surface plasmons transmit very weakly.

In Fig. 4.10 it is shown the transmission spectrum of hole arrays for some common metals. The important observation is that the peaks of the curves correspond to the excitation wavelengths of SPPs at the metal-air and metal-glass interfaces. The peaks of lowest transmittance concern modes in the glass substrate.



(a)



(b)

Figure 4.10: Transmission spectra of hole arrays on misc metallic films.

[9] Primarily their electronic structure, as all the other properties come directly or indirectly from it.

From Fig. 4.10 it is apparent that the influence has the type of metal on the transmittance is significant. For Au, Ag, Cu and other such metals the transmittance is greater than that of Ni and Cr, greater even than that in the PEC case [206]. The simulation results are in good agreement with the experimental measurements [195].

The difference in performance of miscelaneous metals in the transmission through hole arrays, is attributed to their different optical properties and response, mainly the plasmonic frequency, absorption and skin depth. These factors differentiate the metal's capability to excite and sustain SPPs. Regarding Fig. 4.10, Au, Ag and Cu exhibit a quite large penetration depth for the electromagnetic field; this enlarges the effective area of holes facilitating the tunneling[10]. In contrary, Ni and Cr absorb quite a lot, therefore weaken significantly the transmission resonance effect. Similar is the interpretation for the transmission behavior of other metals too.

Especially for the Al, Fig. 4.10a shows that at short optical wavelengths exhibits a PEC behavior. However, at wavelengths longer than ~ 700 nm absorption dominates and the transmission resonance is eliminated. The same holds for Ni and Cr, but they have quite lower transmittance peaks. For W (not shown here), the transmittance peaks are even lower [206] and the transmission resonance even weaker; its effectiveness to transmission is much worse than Ni and Cr. Moreover, the position of the transmittance peaks, and their FWHM, depend on the metal too [206, 195].

In general, the influence of metal type to the transmission is of the same importance as the geometric characteristics, and the metal to be used must be selected appropriately to the application in mind.

4.4.4 Representing hole arrays by anisotropic media

When the characteristic size dimension of a nanostructure is much smaller than the wavelength of the waves involved, then the nanostructure can be regarded as an effective medium. Then an effective dielectric permittivity can be attributed to the nanostructure, corresponding to it as a whole. The same holds for the magnetic permeability. With these effective constants the wave-propagating properties of the device are determined [113], avoiding the cumbersome handling of its structural details. This technique also applies to hole arrays; in this manner, a hole array on a metallic film can be represented by an anisotropic homogeneous material. In this case, the effective optical properties of the holey film are determined by the geometric parameters of the holes [114].

The technique has been applied for an 1D slit array and for an array with square holes; these two cases will be briefly presented here. The geometry is shown in Fig. 4.11. The incident wave excites many waveguide modes inside the holes; however, due to the condition of the small diameter size limit, only the fundamental mode dominates, whereas the modes of higher order are evanescent [166]. Under this condition, for the slit array shown in Fig. 4.11a, the effective optical constants (relative values) are [114]

$$\varepsilon_x = \frac{d}{a \sin c^2(k_x a/2)} \approx \frac{d}{a}, \tag{4.8a}$$

[10] This phenomenon was already discussed at the end of §4.3.

$$\mu_y = \mu_z = \frac{a \sin c^2(k_x a/2)}{d} \approx \frac{a}{d}, \tag{4.8b}$$

$$\mu_y = 1, \tag{4.8c}$$

where a and d are geometric parameters, shown in Fig. 4.11a.

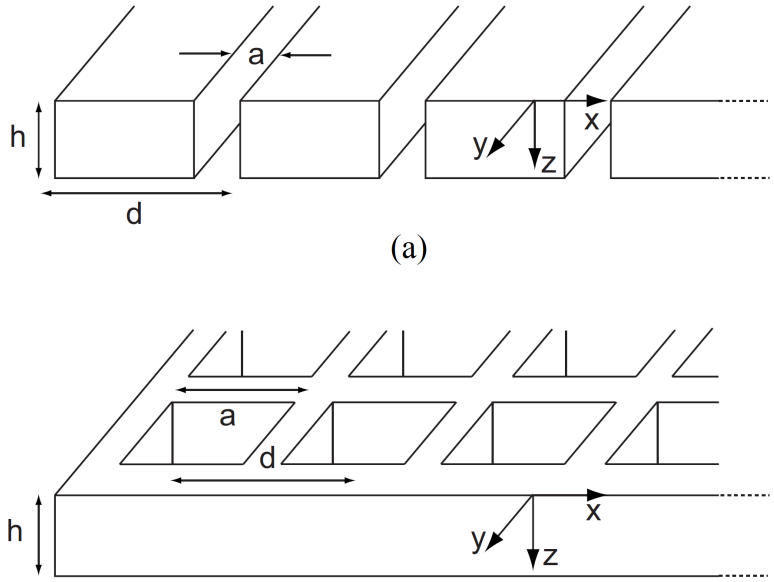


Figure 4.11: Schematics of hole arrays on metallic films.
(a) Array of infinite-length slits. (b) Array of square holes.
The incident wave is p-polarized with parallel momentum k_x .

In the case of the array with the square holes, Fig. 4.11b, the dielectric permittivity and magnetic permeability tensors must be symmetric, that is $\varepsilon_x = \varepsilon_y$, and $\mu_x = \mu_y$. Moreover, the waveguide mode inside the holes exhibits no dispersion for the parallel momentum; this implies the additional requirement $\varepsilon_z = \mu_z = \infty$. Under the condition of small size limit, $d \ll \lambda$, the effective optical constants (relative values) are found to be [114]

$$\varepsilon_x = \varepsilon_y = \frac{d^2\pi^2}{8a^2\varepsilon_h}\left(1 - \frac{\omega_c^2}{\omega^2}\right), \tag{4.9a}$$

$$\mu_x = \mu_y = \frac{8a^2}{d^2\pi^2}\,. \tag{4.9b}$$

In (4.9), ε_h is the permittivity of the dielectric medium inside the hole,
and $\omega_c = \pi c/(a\sqrt{\varepsilon_h})$ the cutoff frequency[11] of the hole regarded as a waveguide.

Summarizing, the metallic holey films of the above forms can be treated as homogeneous metallic media, having the simple effective optical constants given by (4.8) and (4.9) [114]. Adapting appropriately the geometric parameters of the hole arrays, it is possible to create artificial media in a wide range of desired (effective) optical properties. Furthermore, the technique can even be applied to create media with negative refractive index [41]. Although this possibility is still under investigation, studies indicate that this feasible [165].

4.5 Transmission through a hole surrounded by corrugations

In all the above it was mentioned several times the crucial role that SPPs play in the transmission through subwavelength hole arrays. Their contribution is due to the phase-matching of the incident radiation to the SPPs and the tunneling they are subject through the holes. But similar effects also occur for a single isolated hole when surrounded by an array of opaque surface corrugations[12]. Moreover, an appropriate arrangement of corrugations on the exit surface can shape a wave beam with very small deviation angle and be used to control the directionality of the transmitted light. Some such important cases of transmission through single a hole, assisted by corrugations, will be presented below.

4.5.1 Slit surrounded by periodic corrugations

In Fig. 4.12 it is shown the geometry of a slit aperture surrounded by an array of parallel grooves. As a general remark, for apertures allowing a propagating mode, such as an essentially 1D slit structure, where the fundamental TEM mode does not exhibit a cutoff width, the transmission is a very complex process : in this case, the transmission can be modulated via resonances of the fundamental waveguide mode of slit [192], controlled by the thickness[13] of the metal film.

In Fig. 4.13a it is shown the transmission spectrum of a slit type of Fig. 4.12 for a variety number of grooves on the film. The case of a slit without grooves is also included for comparison reasons. For the slit without grooves the transmittance exhibits

[11] ω_c can be considered as the effective plasmon frequency of the anisotropic holey film.

[12] In specific, transmission through a hole on a surface regularly engraved takes place via tunneling, resulting to an approximately exponential dependence of the transmitted intensity on the thickness of the metal film. However, if the thickness is of the order of the skin depth, and the adjacent dielectric media at the front and back interface are the same, then coupling between SPPs at the two surfaces takes place, enabling phase-matching [163].

[13] For this reason, in some studies the wavelength λ is measured in units of the thickness d of the film.

two peaks, quite high. The maxima in this wavelength range are due to Fabry-Pérot resonances[14] inside the space of the single slit [167].

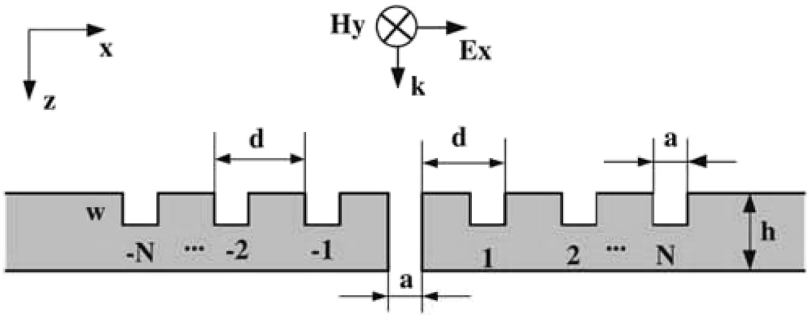
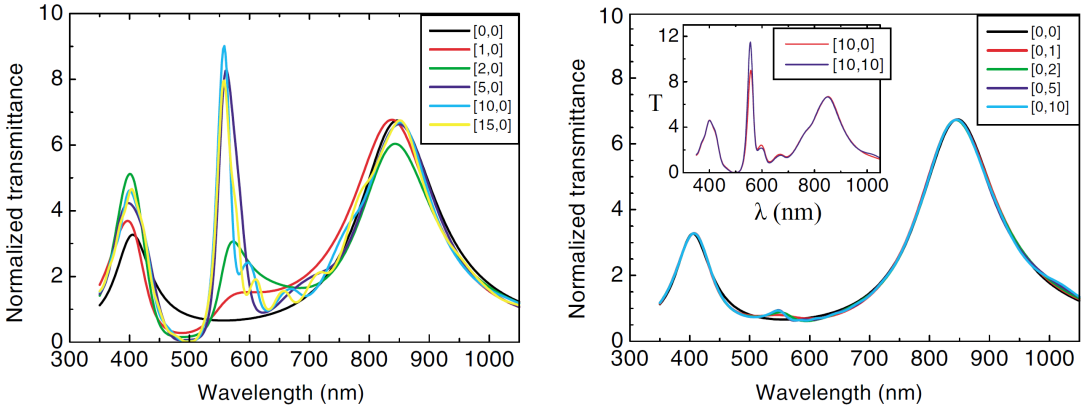


Figure 4.12: Schematic of a single slit surrounded by periodic grooves on the front surface of a metallic film.



(a) (b)

Figure 4.13: Normalized-to-area transmittance $T(\lambda)$, as a function of the number of grooves in (a) front side and (b) back side configurations [74]. Geometrical parameters used in both graphs are $a = 40$ nm, $d = 500$ nm, $h = 350$ nm, and the depth of the grooves $w = 100$ nm. Notation $[NI, NO]$ means $2NI$ grooves in the input surface and $2NO$ grooves in the output surface. In all surfaces, the grooves are located symmetrically around the slit, according to Fig. 4.12. Inset in panel (b) shows $T(\lambda)$ for $[10, 0]$ and $[10, 10]$ patterns. In all cases the incident radiation is p-polarized.

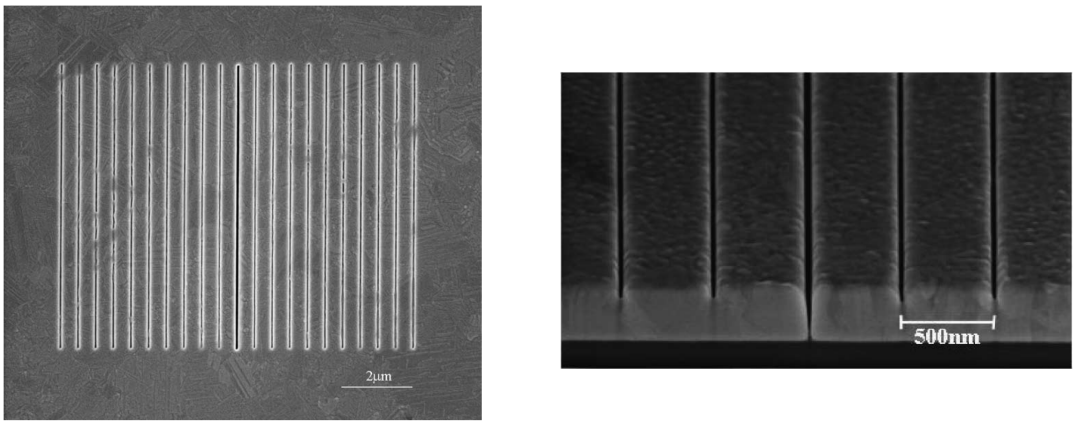
[14] In optics, a *Fabry–Pérot interferometer* or *etalon* is an optical cavity made from two parallel reflecting surfaces (thin mirrors). Optical waves can pass through the optical cavity only when they are in resonance with it. It is named after Charles Fabry and Alfred Perot, who developed the instrument in 1899. *Etalon* comes from the French *étalon*, meaning "measuring gauge" or "standard". (from Wikipedia)

When there are grooves around the slit, additional peaks occur and the transmission is in overall enhanced, Fig. 4.13a. This enhancement is due to coupled cavity modes that are excited because of the grooves [75]. On the front surface, increasing the number of grooves, the existing peaks are getting higher, and a new strong peak also occurs; in general the transmission is improved. For a noteworthy enhancement, a small bunch of grooves (about 10 to 15) is enough. This seems to be a saturation limit; more than this does not further improve the transmission.

The above results concern the front surface, where the wave impinges. A similar patterning only on the back surface, has no effect in the transmission, Fig. 4.13b.

By having grooves on the back surface in addition to the front, the enhancement is practically negligible, as the inset in Fig. 4.13b indicates. Only the new peak becomes a bit higher. The conclusion is that to obtain EOT characteristics from a single slit, a small array of a grooves around the slit, in the front surface, is enough.

At this point an important remark about the polarization must be done. The resonant transmission through a hole is close related to localized SPPs that are excited on corrugated metallic surfaces [167, 97]. The SPPs are naturally p-polarized, thus only the p-polarized component of the impinging wave can benefit from the SPPs and be transmitted resonantly through subwavelength holes and slits; the s-polarized component is suppressed. This limitation can be overcome setting a dielectric layer on the surface of the metallic film; then, dielectric waveguide modes with s-polarization are supported too. If the corrugations are covered by a dielectric layer, then the structure with the slit exhibits the EOT features for both p- and s-polarizations. In this way, a dielectric slab on the metallic film can change profoundly the diffraction and the transmission behavior of an incident plane wave [168, 176].



(a) (b)

Figure 4.14: Real images of a single slit on a metallic film, surrounded by periodic corrugations.
(a) The film with the 1D slit, symmetrically flanked by grooves [77].
(b) Detail of the film cross section (taken using focused-ion-beam milling) [74].

4.5.2 Directional emission using corrugations

As discussed above, corrugations engraved on the back (exiting) surface of the film have no influence to the transmission spectra. Nevertheless, an array of grooves on the back surface narrows the angle of the exiting radiation [176], thus affecting the directionality. In this manner, hole structures with apropriate patterns in both the front and back surfaces, Fig. 4.15, exhibit not only enhanced transmission but also directional control on the emission.

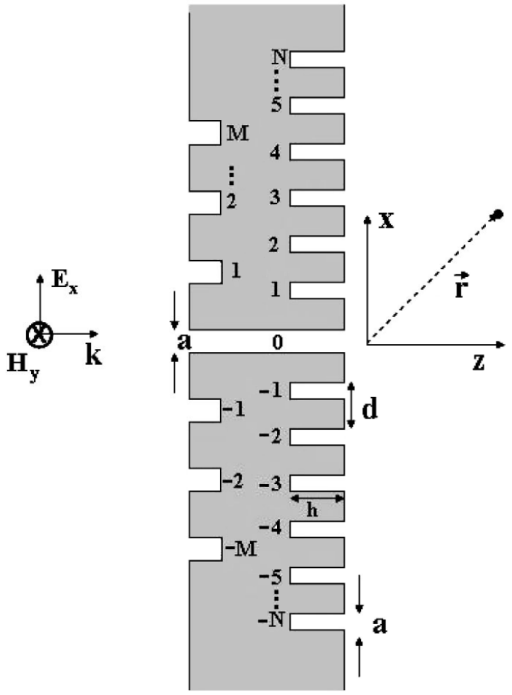
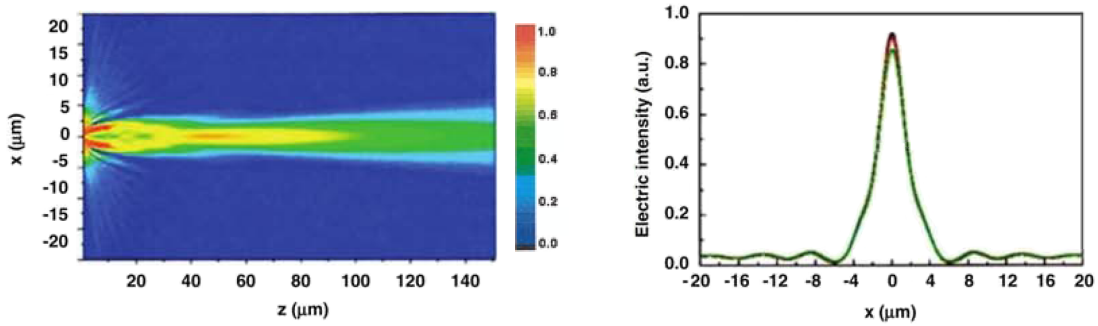


Figure 4.15: Schematic of a single slit surrounded by 2M grooves on the front (entrance) surface and by 2N grooves on the back (exit) surface [75]. The grooves are symmetrically engraved on either side of the slit.

In Fig. 4.16 it is demonstrated the influence has on the directionality of transmitted radiation an array of grooves on the exiting surface of a film [75]. The film with the slit and the grooves has the form of Fig. 4.15, with grooves on both the front and back surfaces. Specifically, in Fig. 4.16a it is shown the intensity profile of the electric field exiting from the slit. There is an evidently collimated beam, with small deviation. The beam also exhibits an elongated focus depth, caused by the collimation. This focusing phenomenon of the electric field takes place in the transitional region between the near- and far-field. The narrow shape of the beam is also shown in Fig. 4.16b, where a variety number of grooves on the exiting surface is tested. In conclusion, patterns on the exiting surface can provide elongated focusing and narrow deviation, giving in this manner quite control on the directionality of the transmitted beam.



(a) (b)

Figure 4.16: Influence of grooves on the exiting surface to the directionality [75]. (a) Electric near-field intensity, (b) cross intensity along $z = 46.3$ μm in (a). The film is of form Fig. 4.15, with geometric parameters $N = 10$, $a = 40$ nm, $h = 83.5$ nm, $d = 50$ nm. The resonance wavelength is 532 nm. In (b) there is a bunch of curves, corresponding to a set of (a, h) parameters.

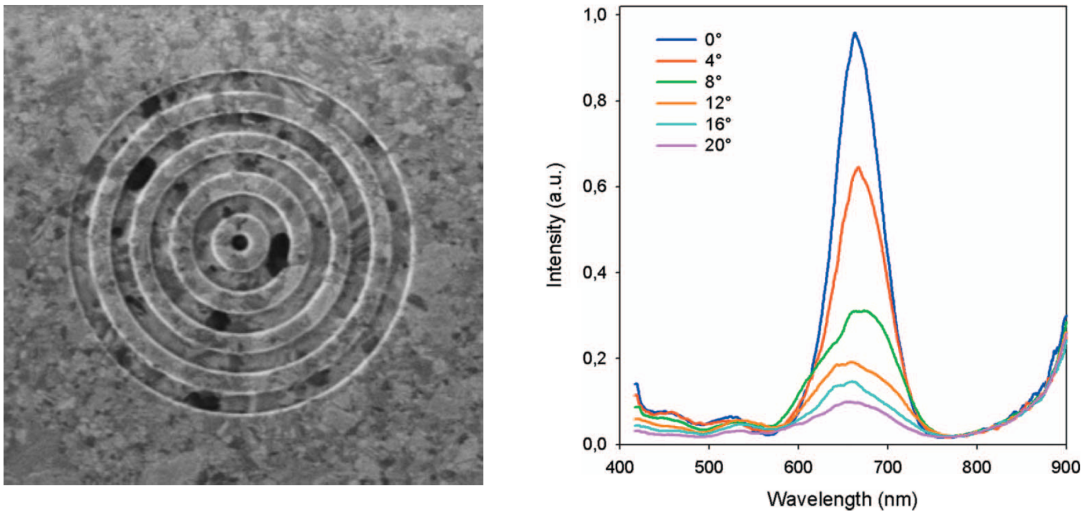
4.5.3 Circular hole surrounded by concentric corrugations

Another interesting case is a circular hole flanked by concentric corrugations[15], on both surfaces of the film. In Fig. 4.17b it is shown the transmission spectrum of such a structure [77], where the focusing of the transmitted radiation it is evident. As in case of the slit discussed above, the resonance wavelength and the transmission maximum are determined by the coupling condition of the pattern on the front (entrance) surface, whereas the width and the directionality of the exiting radiation are controlled by the patterns on the back (exit) surface of the film.

In Fig. 4.17b it is clear that the resonance wavelength is indeed independent from the measure angles, exhibiting only a slight deviation of the exiting radiation. More specific, at FWHM the deviation of the beam is about 5°, a satisfactory result that provides good control on the intensity amplitude. It is reported [77] that the light emitted from the exiting surface at the resonant wavelength is focused on a circular area with radius 1 μm or less. In general, this aperture structure has two noteworthy features [151] :

- The focus length and the shape of the exiting beam are independent of the incident angle. This means that the aperture structure can focus the light from a broad solid angle into an invarying spot.
- The focusing effect occurs in a resonant condition, in which only a narrow band around the resonant wavelength can be focused.

[15] This arrangement is also known as *bull's eye geometry*.



(a) (b)

Figure 4.17: (a) Real image of a bull's eye structure.
(b) Transmission spectrum for a variety of measure angles [77].
A 250 nm diameter cylindrical hole is milled into a 300 nm thick silver film. It is surrounded by five circular trenches with depths of 60 nm and the groove periodicity is 600 nm. The exiting beam is clearly focused. The tail above 800 nm is due to experimental noise.

The transmission spectrum of a hole surrounded by periodic corrugations is determined mainly by the corrugations and the hole depth (or the film thickness)[16]. It was found experimentally that the transmission intensity and the resonance wavelengths have an exponential dependence on the hole depth and the corrugation period [50]. Also, the transmitted radiation exhibits frequency shifting because the corrugations change the effective dielectric permittivity[17]. To this modified effective permittivity corresponds an SPP with wavelength roughly equal to the corrugation period. The SPPs excited by the concentric corrugations dominate versus the waveguide modes arising inside the hole, and contribute to the transmission enhancing much more. A quantity to compare the transmission enhancement caused by the grooves is[18]

$$\eta = \frac{\int I(\theta)\, d\Omega}{\int I_N(\theta)\, d\Omega}, \tag{4.10}$$

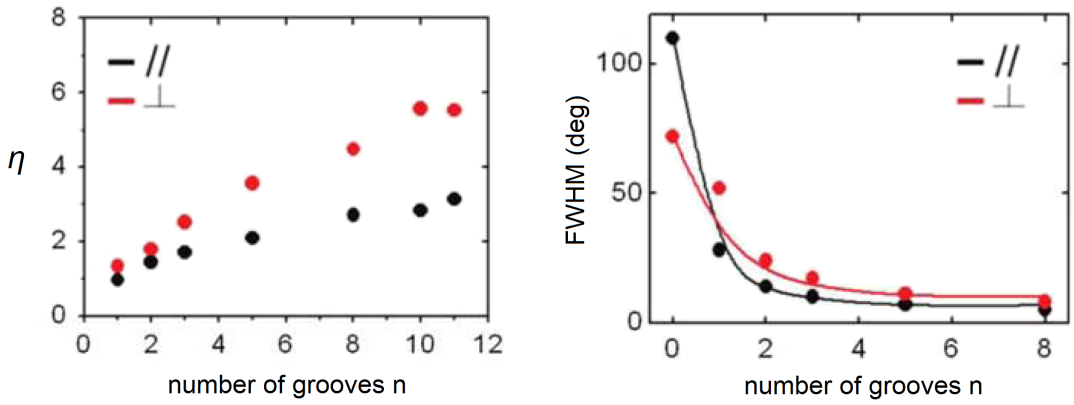
where $I(\theta)$ is the intensity as a function of the polar angle θ and the integration is over a total solid angle Ω. The denominator in (4.10) is simply a normalization factor, where the integral is over the same region and $I_N(\theta)$ is the same as $I(\theta)$ but for the hole without the surrounding corrugations. A name for η is not universally established but it could be called *transmission enhancement coefficient*.

[16] See also F/note 12, p. 94.

[17] Frequency shifting (redshift or blueshift) takes place also in single holes without corrugations (see the discussion of Eq. (4.7), p. 85) and in hole arrays (see Fig. 4.7, p. 89) but it is due to other causes.

[18] Originally defined in [275]. Not to be confused with the usual transmission coefficient defined in (4.2).

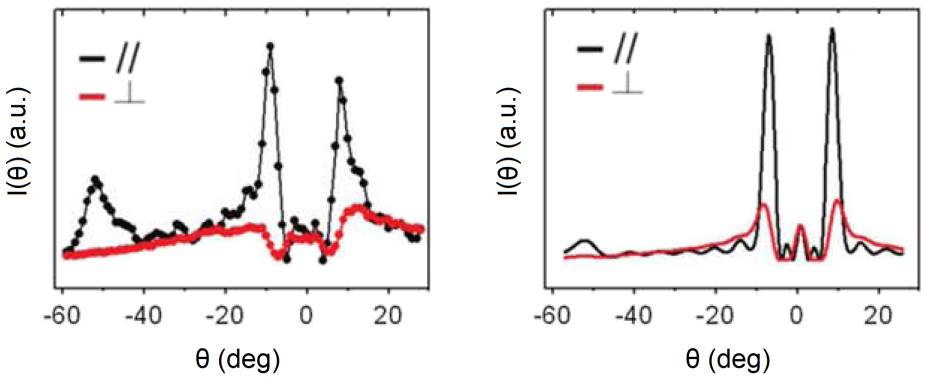
An interpretation of the directionality and the focusing effects of a cirular hole can be given by the Huygens-Fresnel theory; in this frame, the transmitted intensity at the far field comes from the interference of the direct transmitted wave and the scattered wave from the corrugations [50, 275].



(a) (b)

Figure 4.18: (a) Normalized enhancement factor, (b) angle of the FWHM lobe as functions of the number of grooves, in p- and s-polarization for the incident wave.

In Fig. 4.18a,b it is shown η and FWHM as functions of the number of grooves in a bull's eye structure, under p- and s-polarization of the impinging wave [275]. It is evident that η increases gradually with the number of grooves, and even more when the wave is s-polarized. The FWHM drops very rapidly with only a few grooves, that is the beam becomes narrower (more directional), as discussed above.



(a) (b)

Figure 4.19: Far field transmission intensity (normalized) as a function of polar angle, and under p- and s-polarization for the impinging wave. (a) Experimentally measured. (b) Theoretically calculated.

Also, as shown in Fig. 4.19 [275], the intensity distribution $I(\theta)$ in the far field has a complex shape with many lobes, not a single localized spot. It also depends strongly on the polarization : p-polarization gives much higher intensity peaks than s-polarization[19]. The experimental results agree with the Huygens-Fresnel theory, and indicate that the number of grooves and their periodicity, together with the polarization, are crucial for the directionality properties – at least for this hole structure [50].

4.5.4 Arrays of annular holes and other geometries

From all the above it is concluded that the transmission enhancement and the EOT behavior of hole arrays comes mainly from :

- the elongation of cutoff wavelength of the single hole,
- the waveguide modes arising inside the hole,
- the resonant coupling of SPPs between the holes and their elements (entrance-excit of holes, grooves),
- the SPPs arising on the dielectric-metal interface.

The wave modes arising in these procedures are sensitive to the geometric parameters of the structure, like the diameter and depth of the holes, the density (periodicity) of the hole array etc. Additionally, another factor of great importance for the transmission spectrum and EOT is the shape of the hole. This is reasonable because regarding the hole as a tiny waveguide, its cross section defines the regime of the waveguide modes propagating inside it. Apart from the circular and rectangular holes, other less usual shapes can be used for transmittance enhancement and EOT. Such a case is annular (coaxial ring) holes, to be discussed briefly here.

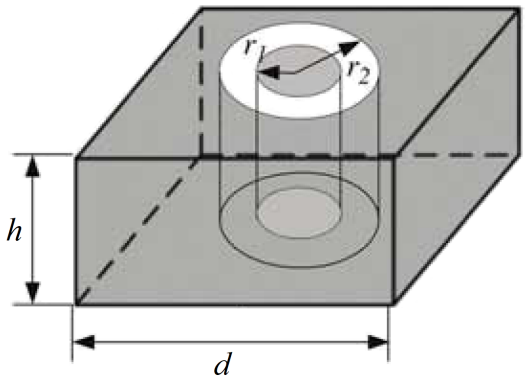


Figure 4.20: Schematic of an annular hole. It can be the cell of an array of such holes.

[19] The reason is that the excited SPPs are naturally p-polarized. See the remark about this issue in p. 96.

In Fig. 4.20 it is shown the geometry of an annular hole. This can be a cell on array of such holes, each one characterized by the inner and outer diameters of the hole, the periodicity of the array and the film thickness (or hole depth). Compared to the holes discussed so far, the transmission mechanism in this hole differs in a crucial way : a special type of surface plasmon is excited on the metal-dielectric interfaces inside the hole, called *cylidrical surface plasmon (CSP)*, which creeps on the walls of the hole. The transmission enhancement and the EOT are mainly due to resonances of CSPs, which act independently from the resonances of the other surface plasmons [275].

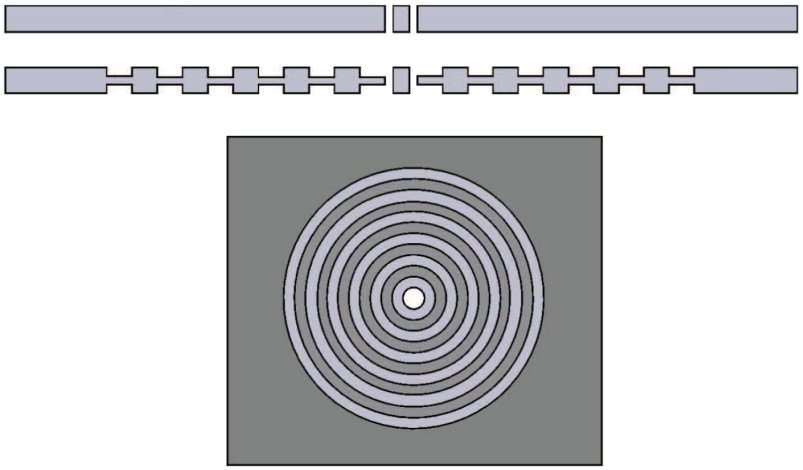


Figure 4.21: Two films with annular holes. The first film is plane. The second one has grooves surrounding the hole in both the front and back surface.

In Figs. 4.21 and 4.22 it is shown two films with a single annular hole, and their transmission spectra [40]. The first film is plane, the second one has circular grooves around the hole on both the front and back surface. As shown in Fig. 4.22, the transmission spectrum has two peaks, one is strong, the other is low but distinct and visible. The first peak, higher and broader, is located at a longer wavelength, at the positions of CSP resonances and is attributed to them. The second peak, lower and narrower, is located at a shorter wavelength. Decreasing of the annulus width, redshifts the peaks to longer wavelength and makes more apparent the transmission enhancement. In general, the decreasing of annulus width makes the coupling of CSPs stronger and more efficient [40]. It has been found that near the infrared wavelengths, the arrays of annular holes give an intensity up to five times that of the same array with circular holes [180, 181, 182]. Also, as shown in Fig. 4.22b, the grooves on the back side broadens the high peak, and redshifts and raises the second one.

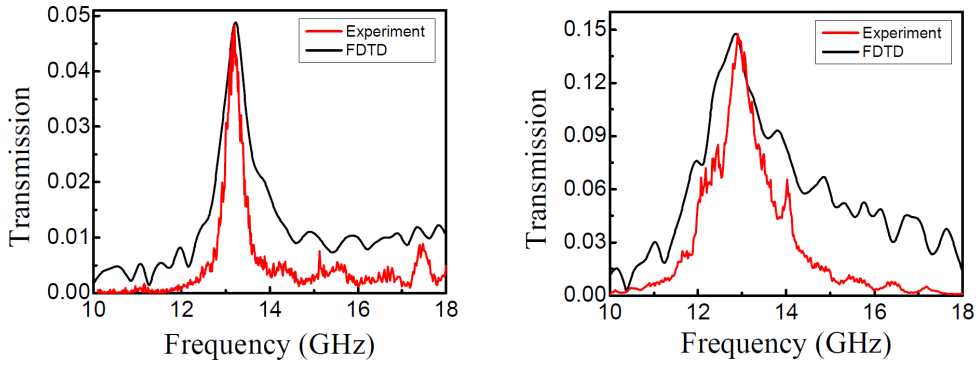


Figure 4.22: The transmission spectra of the two annular holes in Fig. 4.21 [40]. (a) Simple annular hole. (b) Annular hole with surrounding grooves.

Beside the annular holes, a plethora of holes with more unusual shapes were also proposed or tested experimentally, all of them providing enhancement to the transmission intensity. Among them are :

- diamond shaped holes and triangles [123],
- H-shaped holes [236],
- cruciform-shaped holes [43],
- hybrid structure of circular holes and cross-dipole shape [43].

Also, arrays of double holes which are slightly overlapped were investigated. It was found that, compared to usual strucures, at the narrowest spots these structures provide much higher field enhancement. This additional localized enhancement is attributed to nonlinear enhancing effects [274, 148]. All this variety in the available shapes for the holes, together with the geometric parameters of the array, give freedom and versality to the control of the transmission enancement.

4.6 Transmission through a single hole

Concerning a single isolated hole (subwavelength), the transmission can also exhibit EOT behavior, and the factors influencing it are similar as in a hole array; however, the basic component of the process is a bit different. Here the *local surface plasmons (LSPs)* contribute significantly to the transmission, such as the SPPs. This topic will be briefly examined below. It is reminded that Bethe's theory, developed for the light transmission through a subwavelength hole, is inadequate to describe the EOT phenomena, due to two reasons[20] : it considers the film to be PEC and infinitely thin. As discussed in §4.2, these conditions are unrealistic, leading Bethe's theory to wrong predictions for the transmission behavior as really takes place in subwavelength holes.

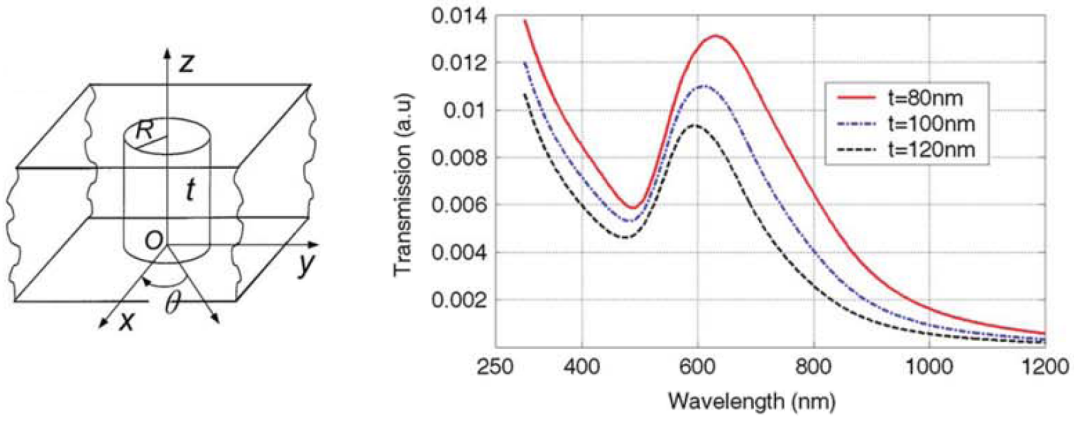
[20] See p. 82.

The presence of LSPs at such a hole, among others induce a noteworthy effect affecting the transmission[21] : the effective diameter of the hole is increased, with all that entails. This phenomenon was already mentioned in §4.3, see Eq. (4.7), and will not be discussed further here.

From the theory of Plasmonics[22] it is expected that the shape and dimensions of the hole will have an important role in the transmission[23]; indeed, they determine the resonance wevelengths in the transmission spectrum. At the rim of the hole occurs a significant field enhancement, resulting in increasing the transmission at the wavelength where the LSP is excited. Below, this behavior will be demonstrated in specific cases and some important results will be presented.

4.6.1 Transmission behavior of a single circular hole

A circular hole is the simplest case for investigating the transmission behavior of single holes. In Fig. 4.23a it is shown the geometry of an indicating case. The film is made of gold and is deposited on a dielectric substrate (here glass), whereas its other surface is free (air). Using FDTD, the transmission spectrum is obtained for some thickness values of the film, Fig. 4.23b.



(a) (b)

Figure 4.23: (reprinted from [151]). (a) Geometry of a single hole in a metal film. (b) Transmission spectrum for such a hole for misc values of film thickness. The diameter of the holes is 200 nm, and the film is gold, deposited on a glass substrate. FDTD was used.

[21] Also, in the case of a circular hole on a metal film with by a free-electron dielectric function similar to that of Drude-Sommerfeld model, it has been found theoretically that below the plasma frequency a propagating mode exists, even for a hole arbitrary small [220, 264]. The influence of this mode to the transmission through subwavelength circular holes remains to be investigated experimentally.

[22] More specific for metal nanoparticles and nanovoids, see [163].

[23] It is reminded that the shape and size of the hole are also important for the transmission spectrum in a hole array, see the discussion of Fig. 4.8 in p. 90.

As shown in the results, in contrast to a hole array, the transmission exhibits only one a peak[24]. Increasing the film thickness, the peak slightly blueshifts and the transmission in general decreases.

The electric field distribution on the surface of the film, and around the hole too, is in accordance with the polarization of the incident field; the cyclic symmetry of the hole is immaterial to this. On the rim of the hole the electric field exhibits two strong peaks[25]. These peaks are caused by opposite charges accumulated on the rim due to the polarization of the incident field, Fig. 4.24. At the edge of the hole the incident field excites an LSP which attenuates rapidly away from the hole [51, 194] and plays important role in the transmission. The electric dipole due to the charges on the edge also radiates (but weaker) and contributes to the local enhancing of the electromagnetic field [42, 191].

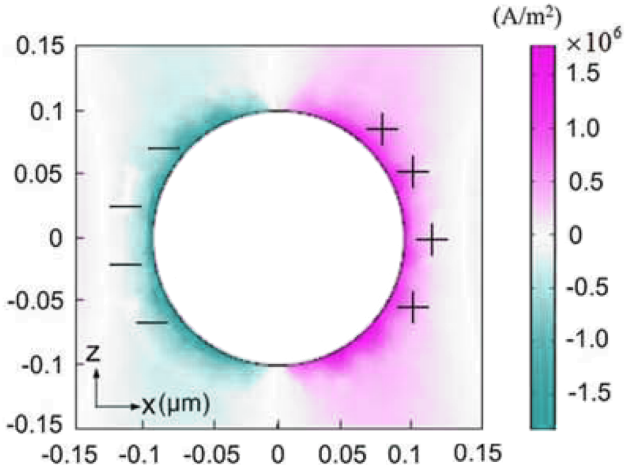


Figure 4.24: Density plot of charge at the rim of the hole. These accumulated charges induce two strong peaks of the electric field on the rim.

In Fig. 4.25a it is shown the electric field distribution, time-averaged, on a plane above the surface of a metal film with a hole. The profile of the field in a section with the hole is displayed in Fig. 4.25b; for comparison, the field is also shown in the same section but without the hole, Fig. 4.25c. Details for the simulation are in [42]. The peaks of the field intensity, shown as concentric rings around the hole, roughly agree with the periodicity of an SPPs excited on the surface. These fringes are caused by constructive and destructive interference of the SPPs with the directly incident field on the surface. In the case of the film without hole, the rings also exist but weaker, and are caused by SPPs excited on the surface [42].

[24] See also the comparative Fig. 4.4 in p. 86.

[25] Sometimes this is known as the *edge effect*.

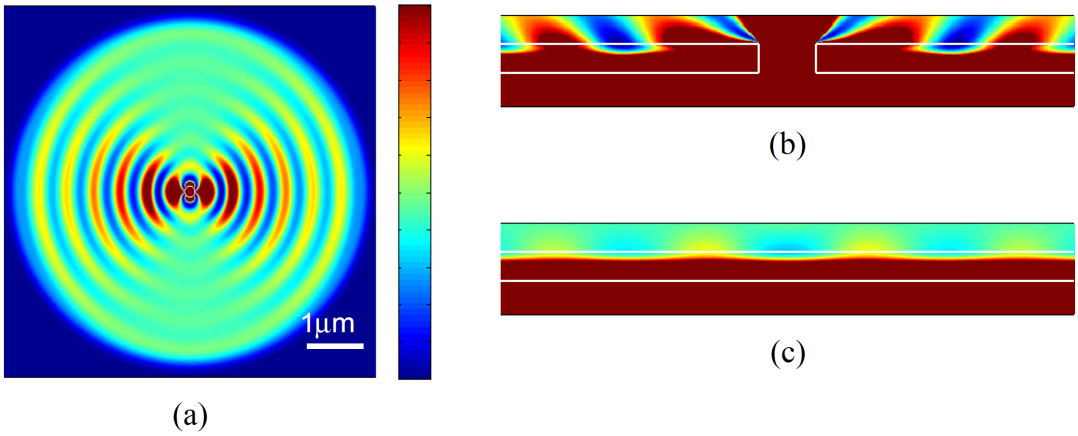


Figure 4.25: (reprinted from [40]). (a) Time-averaged electric field for an isolated nanohole, on the xy plane, $z = 4$ nm. (b) Electric field on xz plane, $y = 0$. (c) The same as (b) but without a hole.

As mentioned earlier[26], the cutoff wavelength λ_c of a single hole in a metal film can differ significantly from the case of a hole on a PEC. The reason is that the skin depth is nonlegligible for some metals in the visible spectrum [51] causing the effective diameter of the hole to increase. Increasing the film thickness (hole depth), the intensity of the peaks at the resonant wavelengths decreases rapidly. Also, beyond λ_c the transmission decays exponentially as the operating wavelength increases[27].

4.6.2 Transmission behavior of a single rectangular hole

The behavior of a rectangular single hole is somehow different from a circular one; now due to the lack of circular symmetry, the role of polarization of the incident light to the transmission is important. Here will be discussed an indicative case.

In Fig. 4.26 it is shown the geometry of a single rectangular hole on a metal film, illuminating by a p-polarized wave [76]. Fig. 4.27a displays the transmmission spectra for a variety of ratios a_y/a_x of the hole sides, the film thickness being constant to $a_y/3$ in all the trials. The intensity is normalized to the hole's area, the wavelength is measured in units of the cutoff wavelength. The azithum angle of the polarization[28] is measured with respect to the y-axis of the structure.

As it is seen, for $a_y/a_x < 1$, and for a square hole, the transmission at resonance is vely low. Increasing the a_y/a_x, the transmission peak increases and becomes narrower. Above the cutoff wavelength the transmission decreases very quickly because the wave inside the hole becomes evanescent and decays very rapidly [76]. In the inset of Fig. 4.27a the transmission spectrum for a square and a circular hole are also presented for comparison; this graph agrees quite a lot with the graph of Fig. 4.2b obtained

[26] See the discussion of Eq. (4.7) in p. 85.

[27] See also the discussion of Fig. 4.2b in p. 82.

[28] It is noted that θ is the incident angle of the wave, not the azimuth angle of the polarization.

by Bethe's theory.

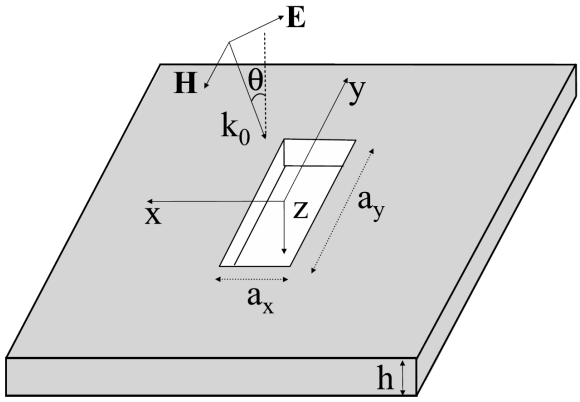
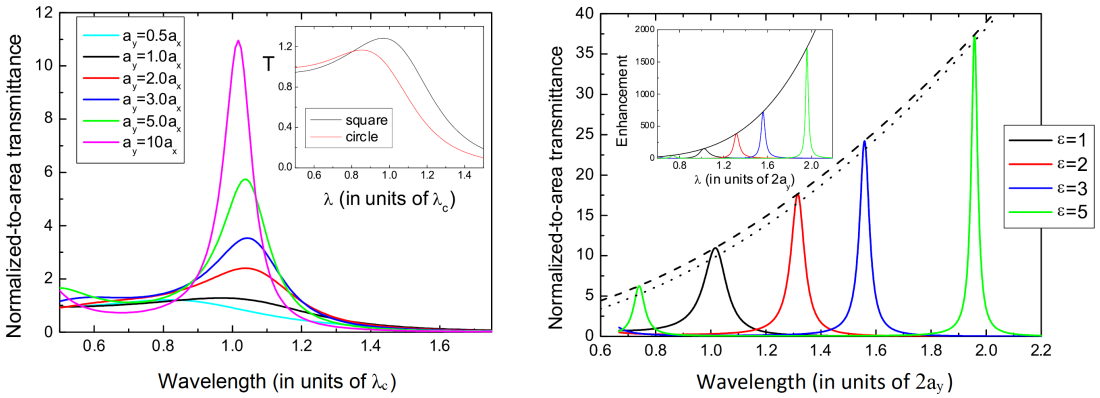


Figure 4.26: Geometry of a rectangular hole on a metal film. A p-polarized wave impinges on the hole.



(a) (b)

Figure 4.27: (a) Transmission spectra for a variety of ratios a_y/a_x of the hole in Fig. 4.26. For comparison, the inset shows the transmission for a rectangular and a cirular hole.
(b) Transmission spectra of a rectangular hole with $a_y/a_x = 10$ and different values of permittivity inside the hole. The inset displays the intensity enhancement of the electric field for these cases.

A rectangular hole exhibits two resonance modes : a longintudinal and a transversal mode, at the wavelength with the peak of lower and higher energy respectively [51]. These resonant peaks are attributed to the excitation of LSPs at the rim of the hole, as in the case of the circular hole discussed previously.

It is remarkable that decreasing the side of a rectangular hole, the normalized peak intesity increases – in contrast to what expected. The same happens for a circular hole too. This counter-intuitive effect indicates that the space of the hole plays a crucial role

in tunneling or coupling of the evanescent waves between the entrance and exiting sides of the hole.

In the subwavelength limit, i.e. when a_x, $a_y \ll \lambda$, the maximum transmission at the resonance wavelength λ_r, for normal incidence, equals approximately to [76]

$$T_r \approx \frac{3}{4\pi}\frac{\lambda_r^2}{a_x a_y}, \tag{4.11}$$

where λ_r is the resonance wavelength of the rectangular hole. Although (4.11) derived for rectangular holes, the same expression holds for circular holes [147], with the term $a_x a_y$ replaced by the area of the circular hole. For the resonance wavelength at the cutoff wavelength $\lambda_r = 2a_y$, (4.11) becomes approximately [76]

$$T_r \approx \frac{3}{\pi}\frac{a_y}{a_x} \ . \tag{4.12}$$

Eq. (4.12) indicates that at the cutoff wavelength, the maximum value of the transmission is approximately proportional to the side length ratio a_y/a_x . This linear result is in good agreement with the results diplayed in Fig. 4.27b. Thus, for the polarization chosen, and for wavelengths near to the cutoff wavelength, the total transmitted light from a rectangular hole depends on the length ratio of the sides.

Also, the propagation constant of the fundamantal TE mode is found to be [76, 77]

$$k_z = \sqrt{\varepsilon_d\, k^2 - \left(\frac{\pi}{a_y}\right)^2}, \tag{4.13}$$

where ε_d is the permittivity of the surrounding medium and $k = 2\pi/\lambda$ is the wavenumber of the incident wave. Eq. (4.13) indicates that the resonance wavelength depends on the permittivity ε_d of the surrounding medium; this results to an approximate relation between the transmission wavelength and ε_d. Therefore, the transmittance at resonance is estimated by Eq. (4.12), but filling the hole with a different dielectric the spectral position of the resonant wavelengths can be shifted according to the relation [76]

$$\lambda_c = 2a_y\sqrt{\varepsilon_d}\,. \tag{4.14}$$

Cosequently, maintaining fixed the ratio a_y/a_x, Eqs. (4.13) and (4.14) imply that the maximum transmission can be increased further by filling the hole with a medium with appropriate dielectric constant, and this will also increase the cutoff wavelength. Such an example is shown in Fig. 4.27b. The side length ratio is $a_y/a_x = 10$ and the thickness is $h = a_y/3$; as evidently seen, increasing the permittivity of the filling media, the transmission and the correpsonding resonance wavelengths are increased. The technique to fill the hole with different media to increase the transmission is valid and can be used for circular holes too.

In the above discussion concerning the transmission through a rectangle hole with varying side ratios a_y/a_x, it is important to point out that the study was done arithmetically and the metal film was modeled as a PEC. Therefore, in contrast to experimental

works like [51], excitation of LSPs along the rim of the hole does not take place, they are excluded by the boundary conditions. The effects and the enhancement observed in the transmission are due to a resonance which however it is reported to be not of surface plasmon nature [163, 76].

4.7 General conclusions for the design of EOT structures

The EOT phenomenon was reported and studied for first time in 1998 by T. W. Ebbesen [55]. Before Ebbesen's landmark paper, subwavelength apertures were regarded to have unavoidably low transmission and strong diffraction. A typical application of subwalength apertures, in which this defect is very troublesome, is the Near-Field Scanning Optical Microscope (NSOM). In NSOMs a subwalength aperture provides the required resolution but with the price of very low signal intensity[29]. The newcome understanding that EM fields can be enhanced strongly at the holes in the metal film, revised and opened new roads for the applications of the subwalength holes. The high performance[30] provided by such structures, together with the capability to adjust their critical properties by sculpting the metal surface, increased furthermore the interest for a wide study of the EOT phenomenon and its potential applications. The main EOT-related disciplines that have been studied the most, are various stand-alone photonic devices and molecular spectroscopy and detection. However, to construct an aperture structure for a given EOT application, many considerations must be taken into account for its structural parameters and the materials, so that the aperture structure to be EOT-optimized for the application. As largely[31] presented in this chapter, the parameters affecting EOT are quite many, and their effect complicated and in combination between them; therefore, designing an optimal EOT structure is not a trivial task. Next, they are summarized roughly the most important results on EOT research, obtained during the last two decades, that have significant implications on possible applications of the EOT phenomenon.

Metal films with holes are robust and can be constructed easily by standard techniques such as focused-ion beam lithography. As the SPPs play a crucial role in EOT, at the operating wavelength the necessary condition must hold for their presence; that is, $\varepsilon_r^{m\prime} < 0$ and $|\varepsilon_r^{m\prime}| > |\varepsilon_r^{i\prime}|$, where $\varepsilon_r^{m\prime}$ and $\varepsilon_r^{i\prime}$ is the real part of the dielectric constant of the metal and the dielectric material in contact with the metal respectively. Furthermore, the imaginary part of the SPP wavenumber[32] must be the smallest possible, so to minimize the damping of SPP by absorption.

---

[29] Bethe's theory is used to explain this low performance [23].

[30] Meaning high transmission and practically zero diffraction.

[31] but even not exhaustively !

[32] For a plane surface it is given by the relation

$$\tilde{k}_{SPP} = \frac{\omega}{c}\sqrt{\frac{\tilde{\varepsilon}_r^i\,\tilde{\varepsilon}_r^m}{\tilde{\varepsilon}_r^i + \tilde{\varepsilon}_r^m}}\ ,$$

where the tilde indicates complex values.

According to their electronic structure, the misc metals support SPPs in different regions of the EM spectrum. For example, Au is suitable only for wavelengths longer than about 550 nm[33], while Ag can be used in VIS and near-IR region of the spectrum[34]. For the UV, Al is a good choice. The roughness of the metal surface is also important; the surface must be as smooth as possible to minimize scattering by the anomalies. The film must be opaque at the operation wavelength; practically, its thickness must be on the order of ten skin depths.

Adjusting the resonant wavelengths (and hence the transmission intensity) can be achieved quite easily by varying the geometric parameters of the structure. In specific, for hole arrays the periodicity, the ratio of film thickness to the holes diameter (Fig. 4.6), the aspect ratio of hole dimensions (Fig. 4.7), and the area of the holes (Fig. 4.8); similarly, for single isolated holes varying the film thickenss (Fig. 4.23), the aspect ratio of hole dimensions and the dielectric inside the hole (Fig. 4.27), as discussed in previous sections.

Generally, if high absolute transmission is required, larger holes should be used – but with the price the resonance will become broader. For sensing or nonlinear effects, high surface filled intensities are required; then the dimensions of the hole relative to the array should be near or below the cutoff of the aperture. Holes like slits can induce Fabry-Perot resonances, enriching the spectrum of the transmission.

Surrounding single holes by periodic grooves is an additional way to enhance and control the transmission. To obtain optimal transmission, besides the aforementioned geometric characterists of the holes, the width, depth or height, and the number of the grooves should be adjusted too. Also, the distance of the periodic grooves from the aperture will influence the outgoing wave through a phase shift.

The metallic film supports SPPs on its both sides, hence the transmission becomes maximum when the SPPs on the two sides are in resonance and the energies of these SPPs coincide. This happens when the refractive indices of the dielectric medium covering the metal are the same on both sides. Simultaneously, the SPPs can couple with the holes in the array, inducing new mode energies and broadening the wavelength operation region. Increasing the film thickness (hence the hole depth) decreases this coupling. To improve the adherence of the SPPs on the metallic film, a thin layer of another metal (usually Cr or Ti) can be paved between the film and its dielectric substrate. In such a structure, if the binding metal has a high imaginary dielectric constant, the SPPs can be entirely damped on this interface.

The finite size of the hole array can also affect the intensity and the peak width of the transmission. For optimal transmission, the array must be definitively larger than the propagation length of the SPPs creeping on the array. This SPP propagation length is much smaller than on a flat film; for example, in the VIS it is smaller about one order of magnitude.

In many cases, it is important to have directionality in the outgoing wave; this can be achieved using a single isolated hole with grooves around it, on the front surface (and maybe on the back one) of the film. Sometimes, for example to characterize the optical properties of a given sample, it is usefull to use a collimated beam on the front surface. This results in a wave with a well-defined wavevector, facilitating the interpretation of

[33] i.e., longer than its interband transitions.

[34] VIS, IR and UV denote the visible, infrared and ultraviolet region of the EM spectrum repsectively.

the transmission when the spectrum or the dispersion curve of the structure is determined.

4.8 A synopsis on the interpretation of EOT

For reference purposes it is useful to give the interpretation of EOT as can be summarized from the phenomena discussed in this chapter [164]. The result of EOT is the enhancement of light transmission through an array of subwavelength holes in a metallic film. Impinging on the film, the incident wave is scattered; its scattering on the hole boundaries creates waves in all directions around, and mainly SPPs along the front surface of the film. When the periodicity of the hole array matches to the resonance condition of SPP excitation, (i.e., it is in agreement with the grating equation), the SPP wave is enhanced significantly; this creates a local field enhancement on the metal surface. This enhancement induces a resonant excitation of the fundamental waveguide mode inside each hole, because the transmission coefficient for its excitation has a resonance that corresponds to the SPP on the interface[35]. At normal incidence, the SPP forms a standing wave on the front interface; then, from across the border of each hole, the electromagnetic energy flows towards the hole, somehow as if the incident power is gathered inside the hole, Fig. 4.28. Besides the flowing SPP, in each hole a radiated field is generated; a part of it propagates along grazing directions to the surface, and can reach the nearby holes and enhances the field there [142, 154].

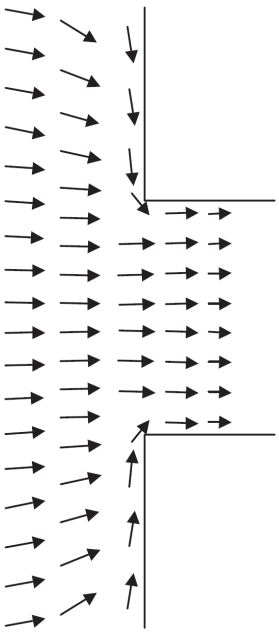


Figure 4.28: Flow of the Poynting vector on the front interface and in the vicinity of the hole entrance, in metal film [164].

[35] Note here that the propagation constant of the SPP is modified by the interaction.

The basis of the phenomenon is that the fundamental mode is resonantly enhanced in this process; this is the reason the field transmitted by the mode is not negligible at the exit side, even if the hole has small cross-section and an evanescent mode is expected from it. At the exit side, the mode excites both an SPP and a radiated field, and the radiation from each hole interferes to form the transmitted wave of the main (zeroth) diffracted order. If the structure is symmetrical (i.e., both sides have identical substrate, cladding, and corrugations), the resonant conditions for excitation of the SPP on the entry interface are the same as for a resonant emission of the SPP on the exit interface into the zeroth transmission order; this enhances the transmission even more.

In the case that the periodicity is not suitable to excite the SPP on the entry side, then, when a SPP is excited on the exit side, another resonance can take place [164]. The waveguide mode inside the holes acts to enhance the tunneling of wave from the entry to the exit surface. When the grating period is suitably chosen, the periodicity can add in phase the SPP generated on each exit hole, thus enhancing the SPP. In addition, the same periodicity also enhances the radiation of the SPP into the substrate[36].

Many other factors can modify the system response and contribute to the quantitative understanding of EOT. For example, the shape of the hole, the type of metal, the film thickness etc. The effect of quite a few of these factors was discussed earlier in this chapter.

[36] This explains the existence of the transmission peak close to 1.4 μm in the famous experiment of Ebbesen et al.

5. The APOTUS Hole[1] Method

5.1 Introduction

The EOT phenomenon through periodic arrays of subwavelength holes, or through a single subwavelength hole with proper configuration (grooves, geometry etc), has received tremendous attention since its discovery in 1998 by Ebbesen and his coworkers [55]. In EOT, the transmitted fraction of the incident light, for certain wavelengths, is generally larger to quite larger than the conventionally expected from the available area of the open holes. This phenomenon, sometimes also been called enhanced or resonant transmission, exhibits similarities to the transmission behavior of frequency selective surfaces for frequencies in the range from the near IR to the microwaves [268, 265]. The large transmission gives the opportunity for many applications˙ for example, in molecular absorption, fluorescence, vibrational spectroscopy (IR and Raman[2]), photonic devices etc [77], some of them being very intriguing. Furthermore, the underlying mechanism of the EOT is complicated and deserves scientific attention for itself.

However, despite the great improvement that EOT brings to the transmission, with whatever this implies for practical applications, there is a plethora of important applications that require light transmission and concentration in a magnitude completely not achievable with EOT or other conventional methods. These are mainly photonic applications relying on nanofocusing, such as :

- all-optical data writing/storage,
- heat-assisted magnetic recording (HAMR),
- near-field scanning optical or thermal nanoscopy,
- nanoimaging, spectroscopy,
- thermal scanning probe lithography,
- nanoscale thermometry,

and others. In all these cases, it is required to focus with high efficiency $\sim$100 μW of energy to a $\sim$10 nm (or less) spot on a planar surface [163, 179, 249, 24, 25]. This is an extremely large light intensity, hundreds of millions of times larger than, for example, the intensity of sunlight on the surface of the earth. This power density is $\sim$1000x the throughput of gold-coated tapered optical fibers used in Near-field Scanning Optical

[1] **APOTUS Hole : Almost Perfect Optical Tansmission Through Unstructured Single Hole**

[2] The inelastic scattering of light in and around the visible region is known either as *Raman scattering*, or, when the interaction is with acoustic waves, as *Brillouin scattering*.

Microscopes (NSOMs), which is the incumbent technology allowing the focus of light on the nanoscale [163, 179]. Typical optical transmission efficiencies of NSOM probe tips are between 10^{-4} - 10^{-5}, while the minimum optical throughput efficiency required for commercial applications to be powered by inexpensive 10 mW diode lasers is $\sim$1%. Conventionally, focusing of light is performed in the far-field with lenses. The shortest wavelength light source (among those that could potentially be mass produced), a 400 nm laser diode, offers a $\sim$200 nm minimum diameter spot in the far-field in air (with an infinite aperture). To reach a $\sim$10 nm (or less) spot in the far-field, it would require an index of refraction larger than 7, which is not available with typical natural materials. A similar limitation is usually found with guided modes in dielectric waveguides, with them too yielding a spot size much larger than the required $\sim$10 nm (or less) spot size [163, 179, 249, 24, 25].

Nevertheless, the above apparently fundamental limitations can be overcome by far using an idea surprisingly simple, and quite easy to implement in practice – at least in comparison with the so far used methods. The idea in its essence is to direct the light to a hole with the required dimensions for focusing, just in front of the desired spot, without permitting back-propagation or backscattering, and force the light to pass through the hole; in this way the hole acts simply as a focusing lens. This technique for nanofocusing has many advantages in comparison with the as far used techniques, and above all exhibites a very large transmittance, in the order of magnitude required for the aforementioned applications. In specific, even in a not ideal case, the transmission coefficient is of the order of unity, i.e., the transmission is almost perfect; this is an incredible achievement for focusing in the deep subdiffraction limit (remember that in conventional methods it is of the order 10^{-4} - 10^{-5} as mentioned above). The technique is reasonably named *APOTUS Hole Method* or *APOTUS-HM*, meaning *A*lmost *P*erfect *O*ptical *T*ransmission *T*hrough *U*nstructured *S*ingle hole.

The key issue in APOTUS-HM is to eliminate back-propagation and backscattering as the wave moves towards the hole and when impinges on it; this means that the wave propagation must be strictly unidirectional. This can be accomplished using a special kind of materials or structures, having the property to break the time reciprocity in the wave propagation [108, 248, 250, 251, 260]. In fact, two types of structures[3] can be used to achieve unidirectionality in wave motion : either a unidirectional but non-topological, or a unidirectional but also truly topological structure. Using appropriate configuration, a special type of wave, namely a surface magnetoplasmon (SMP), propagates along the structure in only one way, reaches the hole and is forced to pass through it and focus on the spot in front of it.

In this chapter the APOTUS-HM will be fully discussed. Although its basic idea is very simple, the physics of its best version relies on the topological materials and their properties, and as a technique that surpasses fundamental limitations of Optics has its own scientific interesting. A relation for the transmission coefficient is derived, which essentially includes all the physics of the technique, and is discussed in detail. Results are presented, indicating its performance in practice. Lastly, its advantages are indicated compared to the conventional transmission techniques, and some issues crucial for its practical implementation are pointed out.

---

[3] As it is explained below, this is not a simple material but rather a combination of two materials, and the wave to be focused propagates on their separating interface.

5.2 Rudiments of Surface Magneto Plasmons (SMPs)

Surface plasmon polaritons (SPPs) are a special type of electromagnetic waves that are confined and propagate along the surface of a conductor, typically a metal or a semiconductor [13]. SPPs are due to the resonant oscillations of the free electrons in the conductor with an incident electromagnetic wave[4]. This resonant oscillation is characterized by a frequency ω_p called *plasma frequency*, which defines the magnitude of the free electrons response to time-varying perturbations [199]. Because SPPs depend on the free electron motions, an external magnetic field will affect the SPPs, due to the Lorentz force which can change the response of the charge carriers. In this case, another characteristic frequency ω_c called *cyclotron frequency* is often used, which is a function of the effective mass of the charge carriers and the intensity of the applied magnetic field [186]. An important consequence of magnetizing the plasmons is that the polarizability of the medium becomes highly anisotropic (that is the permittivity of the conductor becomes a tensor), even though the medium is isotropic when the magnetic field is absent. As a result, SPPs exhibit different properties when they are propagating subject to an magnetic field. In this case, they are reasonably called *surface magnetoplasmons (SMPs)* [34].

SMPs, according to the direction of the applied magnetic field **B**, their wavevector **k** (i.e., the propagation direction of the surface wave), and the direction of the surface, can have three main configurations :

- *perpendicular geometry*,
  in which **B** is perpendicular to both the surface and **k**,
- *Faraday geometry*,
  in which **B** is parallel to the surface and **k**,
- *Voigt geometry*,
  in which **B** is parallel to the surface and perpendicular to **k**.

Compared to the traditional SPPs, SMPs have some remarkable properties. For example, SMPs in perpendicular geometry and Faraday geometry can support pseudo-surface waves; this means they attenuate on only one side of the surface [35, 258]. SMPs in Voigt configuration support the nonreciprocal effect, which means the SMP dispersions are different when they propagate along two opposite directions. Also, unlike SPPs that only have one propagating frequency band which is below the plasma frequency [199], SMPs support two propagating bands (cf. Fig. 5.2).

---

[4] The *plasmon* is a quantum of plasma, i.e., an oscillation quantum of the charge density. The plasmons can be in the bulk of a material or on its surface; in this case they are called *surface plasmons (SPs)*. Plasmons can interact with the electromagnetic waves, whose quantum is the photon. The coupling of a plasmon with a photon is a semiparticle named *polariton*. When this coupling concerns a surface plasmon, this semiparticle is called *surface plasmon polariton (SPP)*.
It is clarified that the SP concerns only the charge oscillations, whereas the SPP concerns both the oscillations and the electromagnetic wave with which they are coupled, as a whole.
Also, roughly speaking, a quasiparticle is a "dressed" particle formed from a bare particle by absorbing correlations from a field [85]. Then, a "bare" particle in a strongly correlated field behaves as if it were a different non-interacting particle in free space. The quasiparticle concept is important since it is one of the few known systematic ways to simplify a quantum-mechanical many-body problem.

In the last years, due to the interest arised by EOT through holes in the nanoscale, a plethora of plasmonic devices have been proposed theoretically and experimentally realized in the visible frequencies [13, 185]. These devices have been mainly concern[5] the subwavelength confinement of electromagnetic (EM) waves [56]. For example, in the metal-insulator-metal structures [140] or slot waveguides [53], EM waves can be confined in a space of the order 0.1 λ. Inspired by these structures, some SMP devices made by metals were proposed [15, 111, 276, 86, 144, 119]. However, all these SMP structures are difficult to realize because they require unreachable magnetic fields. The reason is that in order to observe the effect of an external magnetic field, it is required ω_p, ω_c and the incident angular frequency ω to be comparable. But for a metal in the visible frequencies, ω_p and ω are usually in the order of 10^{16} and 10^{15} Hz respectively. Therefore, a magnetic field with intensity $\sim 10^3$ T is required, which is impossible to realize, even in laboratories. So far, there are two ways to overcome this limitation. The first way is to use ferromagnetic materials in nanostructures, such as Ni and Co [66, 54, 240, 246, 282, 16, 67, 30]. In this manner, the required intensity of the applied magnetic field can be decreased to the scale of μT. But this introduces high losses. The second way is to use semiconductors instead of metals in THz range to decrease both ω_p and ω. The ω_p of a doped semiconductor can be decreased to an order of 10^{13} Hz. Therefore, the required external magnetic field can be less than 2 T [6]. In the last years, such SMP devices, consisting of semiconductors, have been proposed [143, 132, 101, 156, 158].

In general, though less known, SMPs are of the same importance as the SPPs, and have applications equally intriguing. The above notes are enough to appreciate their role in APOTUS-HM. A quite complete review of their theory can be found in [141]. Closing this basic introduction, it is purposeful to discuss the dispersion equation of SMPs of Voigt configuration on a plane surface, as this is the case in APOTUS-HM.

5.2.1 Dispersion of SMPs on a plane surface (Voigt configuration)

In Fig. 5.1 it is shown the geometry of an SMP of Voigt configuration, propagating along the surface of a conductor, along *z*-axis. With this orientation, the electric permittivity of the conductor (metal or semiconductor) is the tensor [102] :

$$\bar{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_{xx} & 0 & \varepsilon_{xz} \\ 0 & \varepsilon_{yy} & 0 \\ -\varepsilon_{xz} & 0 & \varepsilon_{xx} \end{bmatrix} . \tag{5.1}$$

The components of $\bar{\boldsymbol{\varepsilon}}$ in (5.1) are

[5] Especially in comparison with studies of SPPs before the year 2000.

[6] This is the case for the APOTUS-HM considered here.

$$\varepsilon_{xx} = \varepsilon_\infty \left[1 - \frac{\omega_p^2 \, (\omega + i\nu)}{\omega \left[(\omega + i\nu)^2 - \omega_c^2 \right]} \right] , \tag{5.2a}$$

$$\varepsilon_{xz} = -i\varepsilon_\infty \frac{\omega_c \, \omega_p^2}{\omega \left[(\omega + i\nu)^2 - \omega_c^2 \right]} , \tag{5.2b}$$

$$\varepsilon_{yy} = \varepsilon_\infty \left[1 - \frac{\omega_p^2}{\omega \, (\omega + i\nu)} \right] . \tag{5.2c}$$

In the above, ω is the angular operational frequency[7], ω_p is the plasma frequency of the conductor, ε_∞ is the permittivity of infinite frequency[8], and $\omega_c = eB/m^*$ is the cyclotron frequency. Also, B is the intensity of the applied magnetic field, e and m^* are the charge and the effective mass of electron respectively, and ν is the attenuation factor (aka loss factor).

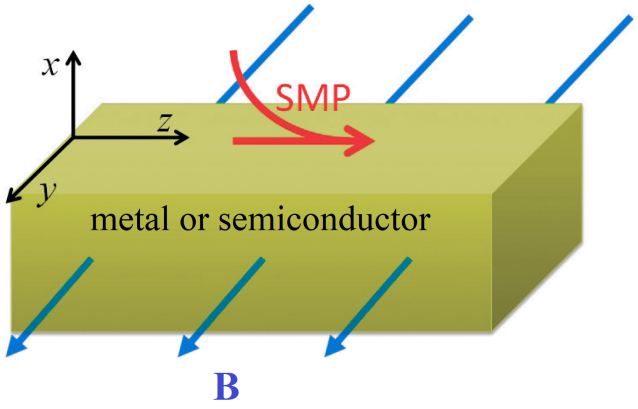


Figure 5.1: Schematic of an SMP of Voigt configuration, propagating on the surface of a conductor.

The wave equation, derived from Maxwell equations, is

$$\nabla \times (\nabla \times \mathbf{E}) - k_0^2 \, \bar{\boldsymbol{\varepsilon}} \cdot \mathbf{E} = \mathbf{0} \ , \tag{5.3}$$

where k_0 is the wavenumber in vacuum, and $\bar{\boldsymbol{\varepsilon}}$ is the electric permittivity tensor[9], in the examined case given by (5.1).

---

[7] That is the frequency of the incident wave exciting the SMP.

[8] ε_∞ is also called *optical dielectric constant*. It is the permittivity in very high (idealy infinite) frequency, and takes values in the range $1 \leq \varepsilon_\infty \leq 10$. The role of ε_∞ is to take account the influence of the background lattice of the material in the dielectric model; specifically, it is the value of ε deduced from the refraction of electromagnetic waves with frequencies high compared to lattice vibrations (phonons). Values of ε_∞ can be found in catalogs with properties of materials. Indicatively, for the isotropic plasma it is $\varepsilon_\infty \approx 1$, for water $\varepsilon_\infty \approx 4.25$ (at 0 - 20 °C), for Au $\varepsilon_\infty \approx 6.5$ - 10, for Ag $\varepsilon_\infty \approx 3.7$ - 4.5 etc.

[9] The factor $\bar{\boldsymbol{\varepsilon}} \cdot \mathbf{E}$ is the dot product of the 2nd order tensor $\bar{\boldsymbol{\varepsilon}}$ with the vector $\mathbf{E}$, resulting in a vector.

Let the material in the region $x > 0$ has permittivity ε_d, and the SMP is TM polarized, as shown in Fig. 5.1. In this configuration, the EM field in the dielectric and in the conductor is

$$\mathbf{E} = \begin{cases} (E_{1x}, 0, E_{1z}) e^{\alpha_0 x} e^{i(\beta z - \omega t)}, & x \geq 0 \\ (E_x, 0, E_z) e^{\alpha x} e^{i(\beta z - \omega t)}, & x < 0 \end{cases} \tag{5.4}$$

where β is the phase constant of the SMP, and α_0, α attenuation constants.
Using (5.4) and (5.1), the wave equation (5.3) has a nontrivial solution only when the these two relations hold :

$$\alpha_0^2 = \beta^2 - \frac{\omega^2}{c^2}\varepsilon_d \ , \tag{5.5}$$

$$\alpha^2 = \beta^2 - \frac{\omega^2}{c^2}\varepsilon_V \ , \tag{5.6}$$

where $\varepsilon_V = \varepsilon_{xx} + \varepsilon_{xz}^2/\varepsilon_{xx}$ is the so called *Voigt dielectric constant*.
Considering the confinement of the mode on the surface, and the continuity of H_y and E_z on the surface, the dispersion equation is eventually obtained :

$$\varepsilon_d \sqrt{\beta^2 - \frac{\omega^2}{c^2}\varepsilon_V} + \varepsilon_V \sqrt{\beta^2 - \frac{\omega^2}{c^2}\varepsilon_d} + i\beta\, \varepsilon_d \frac{\varepsilon_{xz}}{\varepsilon_{xx}} = 0 \, . \tag{5.7}$$

It can be seen immediately that the values $\beta > 0$ and $\beta < 0$ are not equivalent in (5.7), namely the dispersion is non-reciprocal with respect to the propagation direction. Eq. (5.7) can be solved numerically.

In Fig. 5.2 it is shown the dispersion curves of an SMP on a plane surface, as obtained by solving (5.7). The conductor is InSb assuming no losses ($\nu = 0$), with $\omega_c = 0.5\omega_p$, whereas the dielectric is air, $\varepsilon_d = 1$. The rest parameters was taken to be [34] $m^* = 0.014\, m_0$ where m_0 is the free electron mass, $\omega_p = 12.6$ THz and $\varepsilon_\infty = 15.68$. The non-reciprocity of the dispersion is evident in this diagram too. The noticeable feature of the dispersion curve is that for either $\beta > 0$ or $\beta < 0$ the curve has two branches, that is two bands of propagation exist, in contrast to the traditional SPPs. The two cases $\beta > 0$ and $\beta < 0$ are discussed below.

case $\beta > 0$

The lower branch starts from the origin, moves at the right (that is below) of the light curve $\alpha_0 = 0$, and eventually terminates reaching the dispersion curve $\alpha = 0$ of the bulk magnetoplasmons.
The higher branch starts from the curve $\varepsilon_{xx} = 0$ and tends to the asymptotic frequency for non-retarded magnetoplasmons, $\varepsilon_d + \varepsilon_{xx} - i\varepsilon_{xz} = 0$. The phase constant β at which the higher brach starts, is given in reduced form by the relation [102] :

$$\zeta_s^2 = \left(\frac{c\beta}{\omega_p}\right)^2 = \frac{1 + \Omega_c^2}{1 - (\varepsilon_d/\varepsilon_\infty)^2\,(1 + \Omega_c)^2/\Omega_c^2} \ , \tag{5.8}$$

where $\Omega_c = \omega_c/\omega_p$.
For the ζ_s to be positive and finite, it must hold $\varepsilon_d/\varepsilon_\infty > \omega_H/\omega_c$, where $\omega_H = \sqrt{\omega_p^2 + \omega_c^2}$. In the examined case, for InSb and air it is $\Omega_c \geq 0.0064$, hence, the magnetic field must be $B \geq 0.064$ T for the higher band to exist.

case $\beta < 0$

The branches have similar behavior as in the case $\beta > 0$. Here, the lower branch starts from the origin, moves at the right (below) of the light curve $\alpha_0 = 0$, and eventually approaches asymptotically the curve $\varepsilon_d + \varepsilon_{xx} + i\varepsilon_{xz} = 0$.
The higher branch starts from the curve $\varepsilon_{xx} = 1$ and eventually terminates reaching the higher dispersion curve $\alpha = 0$ of the bulk magnetoplasmons.

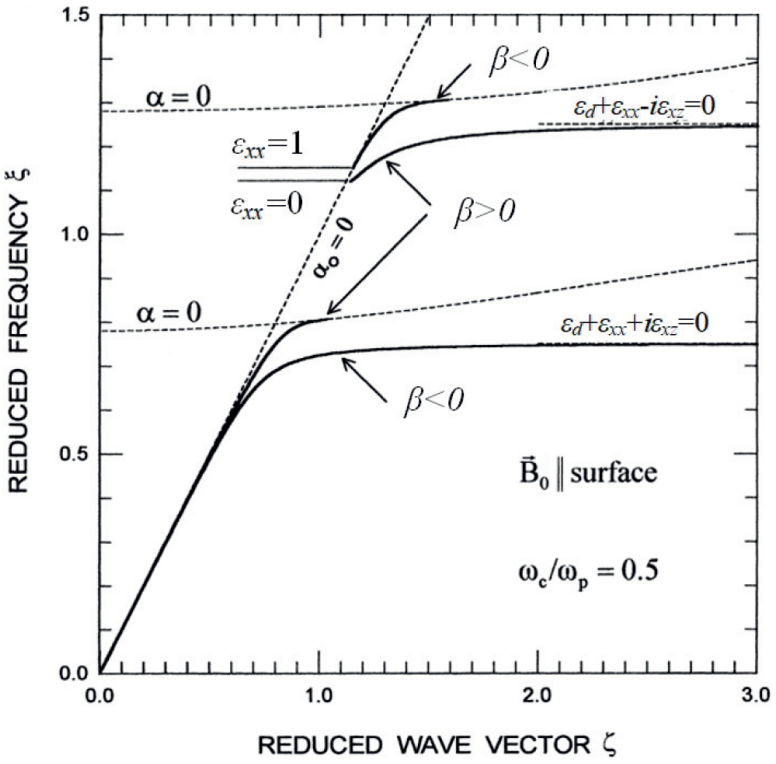


Figure 5.2: (reprinted from [102]). Dispersion diagram of an SMP at the interface of InSb - air in Voigt configuration (solid lines). The $\omega_c = 0.5\omega_p$, $\xi = \omega/\omega_p$, $\zeta = \beta c/\omega_p$ are the normalized angular frequencies and phase constant respectively.

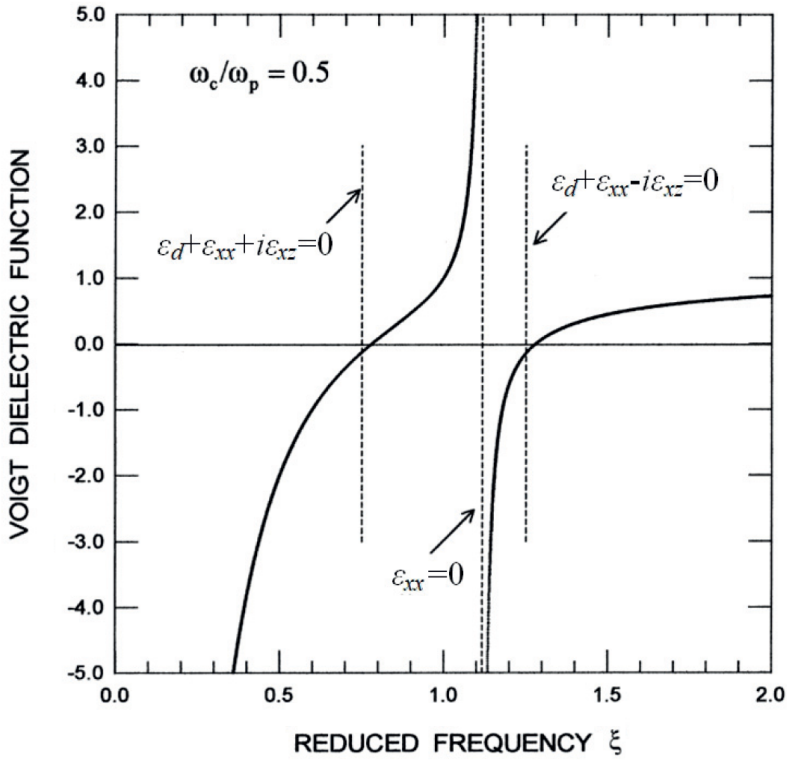


Figure 5.3: (reprinted from [102]). The Voigt dielectric constant ε_V as a function of the normalized angular frequency $\xi = \omega/\omega_p$.

The existence of the two propagating bands (instead of one as in traditional SPPs), can be explained inspecting Fig. 5.3. In this diagram, the Voigt dielectric constant ε_V is sketched as a function of the normalized angular frequency $\xi = \omega/\omega_p$. As it is seen, when $\omega_c = 0.5\omega_p$, there are two curves, and each one has a part where $\varepsilon_V < 0$. But ε_V is the "total" electric permittivity of a structure for SMP propagating. Since SMPs can propagate on the surface of a material with only negative permittivity, SMPs exhibit two propagation bands on a plane surface.

5.3 Configurations and mechanism of APOTUS-HM

As mentioned in §5.1, the APOTUS-HM is based on the unidirectional propagation of the light wave to be focused. Unidirectionality can be achieved with either two configurations : the first (and simpler) is a non-reciprocal but non-topological (NRNT) structure, while the second (and more complicated) is non-reciprocal but also truly topological (NRTT). Concerning the APOTUS-HM, the operating principle of these two structures is practically identical, but they differ in some important characteristics, as it is examined below.

5.3.1 Non-reciprocal, non-topological (NRNT) structure

In Fig. 5.4 it is shown the operational principle of the NRNT structure. The structure is comprised by a two-component material, surrounded by a PEC (here Ag). The one is InSb which is a magnetically biased gyroelectric semiconductor, the other is Si. The one end of the structure is bounded by the PEC, but it has a hole for the wave to pass through and be focused. For computational reasons, the configuration considered here is 2D, thus the materials are extending infinitely along the *y*-axis and the hole is a slit; however, this not an essential limitation : the structure can be 3D (cylindrical) with the cross section being a circle or rectangle. The hole can be as small as required for the applications, typically in the deep subdiffractional range of the wave.

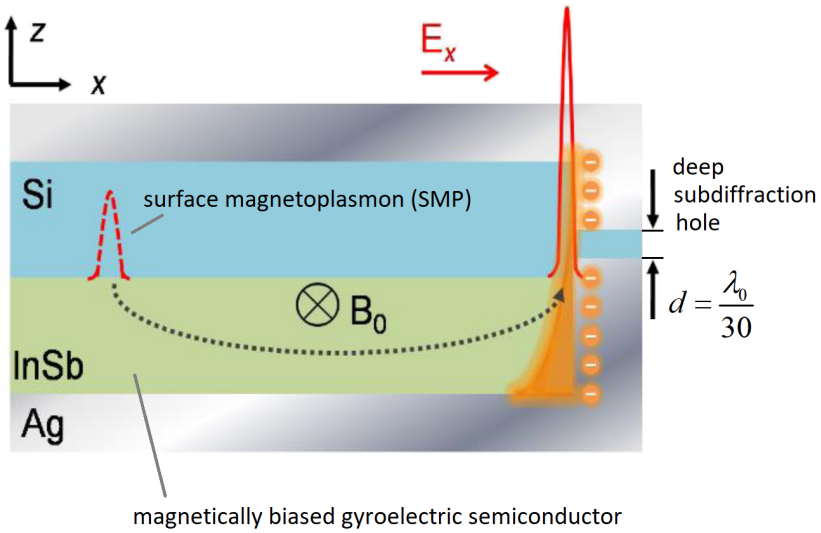


Figure 5.4: Operational principle of the non-reciprocal, non-topological structure for the APOTUS-HM (longitudinal cross section).

Somewhere inside the structure there is a source exciting SPPs. These waves contain the energy that is desired to be transferred and focused on a (subwalength) spot. The

whole structure is subject to an appopriate magnetic field; this means that the SPPs are in fact SMPs and exhibit the properties mentioned in §5.2. In specific, the Voigt geometry[10] is applied for the SMPs, as this configuration supports the nonreciprocal effect. Under the magnetic field, the gyroelectric character of InSb is manifested : its electric permittivity becomes a tensor of the form (5.1) and causes the unidirectional propagation of the wave in the structure.

The SMP travels unidirectionally on the interface between the InSb-Si until it reach the end of the waveguide, where there is the hole with the appropriate dimensions for the focusing. Due to the strictly unidirectional propagation, the wave cannot be reflected and travel backwards; thus, it is obliged to pass through the hole (and then focus in front of it), even if the hole is very small. Therefore, in the frequency band $\Delta\omega$ where the propagation is unidirectional, the transmission coefficient is expected roughly to be unity. This is the basic operational principle ot the APOTUS-HM.

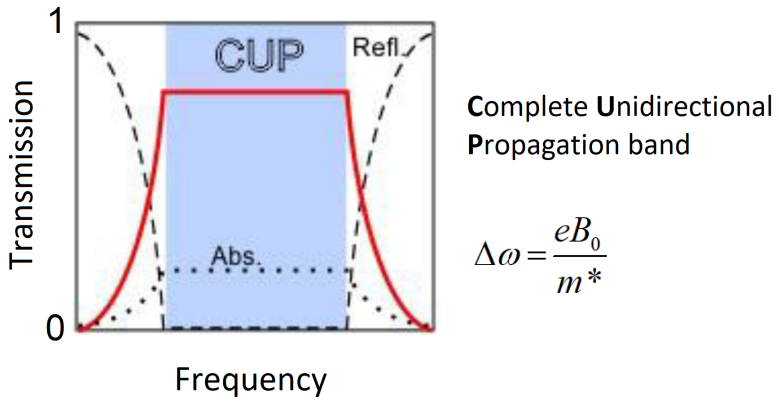


Figure 5.5: Schematic of the Complete Unidirectional Propagation (CUP) band, and the transmission, reflection and absorption coefficients for the the APOTUS-HM structures.

The effectiveness of the APOTUS-HM is founded theoretically on the formula of transmission coefficient through the hole, that is

$$T = \frac{2\gamma_T}{2\gamma_T + \gamma_0}, \tag{5.9}$$

where γ_T is the tunneling rate of the wave[11] through the hole (in the forward direction), and γ_0 is the decay rate of the wave due to dissipative losses. This relation is extracted by Temporal Coupled Mode Theory, see §5.7 and (5.30), and it is very important because it incorporates most of the physics of the transmission phenomena in the device. Next, the physical meaning of (5.9) is thoroughly discussed.

[10] i.e., the magnetic field **B** is parallel to the surface and perpendicular to **k**.

[11] More precisely, the word "wave" means in fact a mode.

In (5.9) two factors are involved, namely γ_0 and γ_T. The inverses of these factors are times; in specific, they are defined as

$\gamma_0 = 1/\tau_{inc}$, where τ_{inc} is the incident wave lifetime,

$\gamma_T = 1/\tau_{tun}$, where τ_{tun} is the tunneling time.

As it seems in (5.9), T depends only on γ_0 and γ_T; and it is immediately evident that when $\gamma_0 \to 0$, that is the material is lossless, or γ_0 is sufficiently small in comparison to γ_T, then the transmission coefficient T tends to unity. Therefore, at least in theory, APOTUS-HM provides the maximum transmission coefficient (i.e., power focusing), no matter how small the hole is compared to the wavelength of the transmitted wave. This is a very important result.

In conclusion, the physical meaning of (5.9) and the factors γ_0 and γ_T is as follows. When the losses of the material are zero or sufficiently small, then the wave near the hole survives for long enough to tunnel and pass through the hole. Therefore, in contrast to other conventional methods, ideally there is no limit in principle for the maximum transmission, no matter how small the hole is.

In practice there are some issues that decrease the transmission from its ideal value, but even so it remains very high compared to the the maximum transmission feasible by other merhods. Some of these issues will be discussed below.

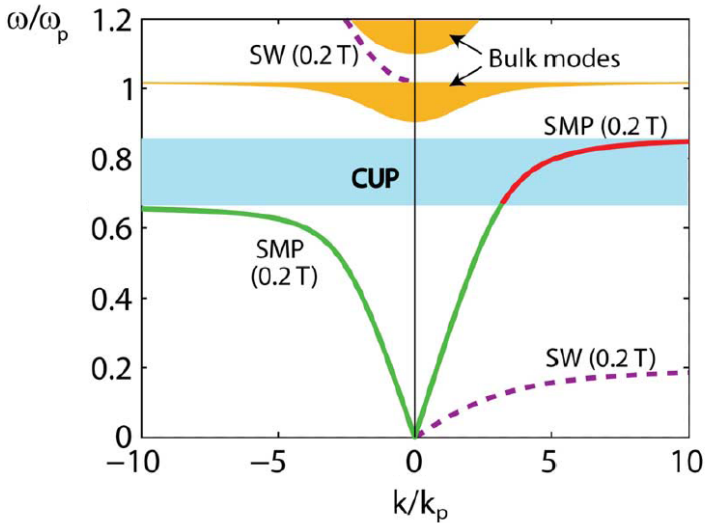


Figure 5.6: Dispersion diagram of the SMPs for the NRNT structure. It is evident the nonreciprocal effect that causes the unidirectional propagation and the CUP band. Some bulk SMP modes are also sketched. The applied magnetic field is 0.2 T.

In Figs. 5.6 and 5.7 it is shown the dispersion diagram for the NRNT structure. As can be seen clearly in Fig. 5.7, when the magnetic field is absent, the dispersion has two symmetric branches, this is the trivial case; but when a magnetic field is applied, then the two branches become asymmetric and a Complete Unidirectional Propagation (CUP) band opens. In the CUP zone the propagation is strictly unidirectional since there is only

the branch corresponding to forward propagation. The wide of CUP band is given by

$$\Delta\omega = \frac{eB_0}{m^*}\,, \tag{5.10}$$

where e is the electron charge, B_0 the magnitude of the applied magnetic field and m^* the effective mass of the electron. This relation can be extracted by the study in [218].

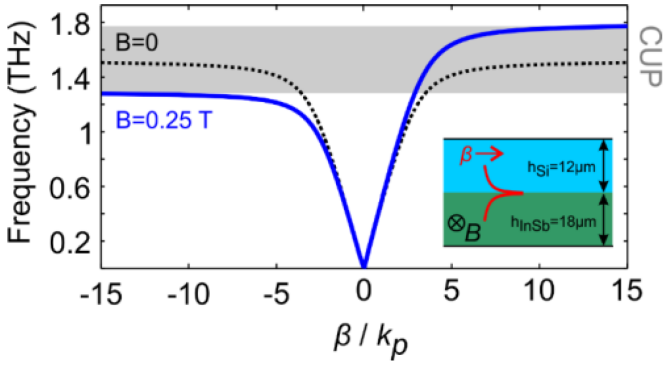


Figure 5.7: Dispersion diagram of the SMPs for the NRNT structure in a real case. The applied magnetic field is 0.25 T.

Note that the use of InSb, besides its gyroelectric property, makes possible to use low values for the applied magnetic field[12] which can be realized in practice quite easily. In general, a magnetic field less than 0.3 T is sufficient to open an operational CUP band.

5.3.2 Non-reciprocal, truly-topological (NRTT) structure

The nonreciprocity is not on its own always sufficient to ensure genuinely unidirectional surface modes. The problem is that in most cases the nonlocal effects (i.e., the spatial dispersion) in the plasmonic material cannot be neglected; such a case is the tight light localization and focusing, taking place in the NRNT structure discussed above. As a result, the CUP band may close, destroying the desired unidirectional propagation and making the operation of the structure unstable. To overcome the detrimental role of nonlocality, truly topological (rather than simply unidirectional) structures are required, such as the NRTT discussed here.

In Fig. 5.8 it is shown a schematic of the NRTT structure for the APOTUS-HM. It is much the same as the NRNT device of Fig. 5.4, but it has two crucial differences :

- the Si layer has been removed, and the semi-infinite InSb layer directly touches the plasmonic (Ag) upper cladding,

[12] See the discussion in p. 116.

- the two-component material is now surrounded by a perfect magnetic conductor (PMC), not a PEC as previously.

The first modification aims to induce topological properties on the structure, thus making the unidirectional propagation stable; the second is related to the geometry of the SMP that is created after replacing the Si layer by the plasmonic material (Ag).

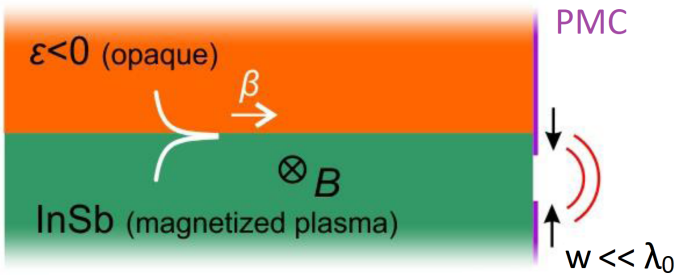


Figure 5.8: Schematic of the non-reciprocal, truly topological (NRTT) structure for the APOTUS-HM (longitudinal cross section).

Thus, in the case of NRTT structure the unidirectionality is due to two mechanisms. The one is the nonreciprocity the InSb exhibits under the magnetic field. The other is the pairing of InSb with the Ag, resulting in a structure with topological properties that, among others, provides robust directionality for the modes propagating in the separating surface of the two media. The unidirectionality imposed to SPPs or SMPs in topological structures is a very complicated phenomenon and its origin is still under investigation (e.g., cf. [65, 72]); here it will be taken for grant, but roughly speaking it is close related with the opening of a bulk-mode bandgap and the Chern number of the topological material[13] (here the pair InSb-Ag).

It is reminded that the Chern number is an integer and can be regarded as the number of windings for a state evolution in the momentum space[14]; in the herein case the state is a mode. Chern number is a topological invariant related to the bulk excitations, and is conserved under continuous deformations that do not close the bandgap[15] [251]. As thoroughly discussed in §1.3, after a full evolution in momentum space, a mode does not in general return back exactly to its initial form; instead, may acquire an additional phase term, the Berry phase, equal to $2\pi C$, where C is the aforementioned Chern number for the given mode. Very crucially, this winding owing to the modal evolution is completely immune to perturbations that are not large enough to destroy the bandgap[16]. As a result, when two media sharing the same bandgap, but having different topological properties, are brought together, the aforementioned bandgap will necessarily close at the separating

[13] Here "material" means the InSb and the Ag together (as a pair), which results in a structure with topological properties. This is the case for the topological materials in general.

[14] See p. 26 and the discussion in p. 39.

[15] See the discussion in §1.4.3 for the bulk-edge correspondence.

[16] since it is (the winding, i.e., the Chern number) a topological invariant and does not change as long as the bandgap does not close.

interface in order to facilitate the change in the topological invariants of the two media; this gives rise to truly topological, unidirectional SMPs at that interface. These waves are immune to even nonlocal plasmonic effects, and their dispersion extends over the full bandgap [251, 65, 72]. Furthermore, the bulk-edge correspondence principle[17] ensures that the total number of unidirectional surface waves at the separating interface of the two media is equal to the difference between the "gap Chern" numbers of the two media[18]. For the herein structure, formed by the interface between a magnetized plasma and an impenetrable material, the difference between the two gap Chern numbers is exactly equal to unity; as a result, exactly one unidirectional surface wave extends across (closes) the bandgap [72], ideal for the requirement of the APOTUS-HM.

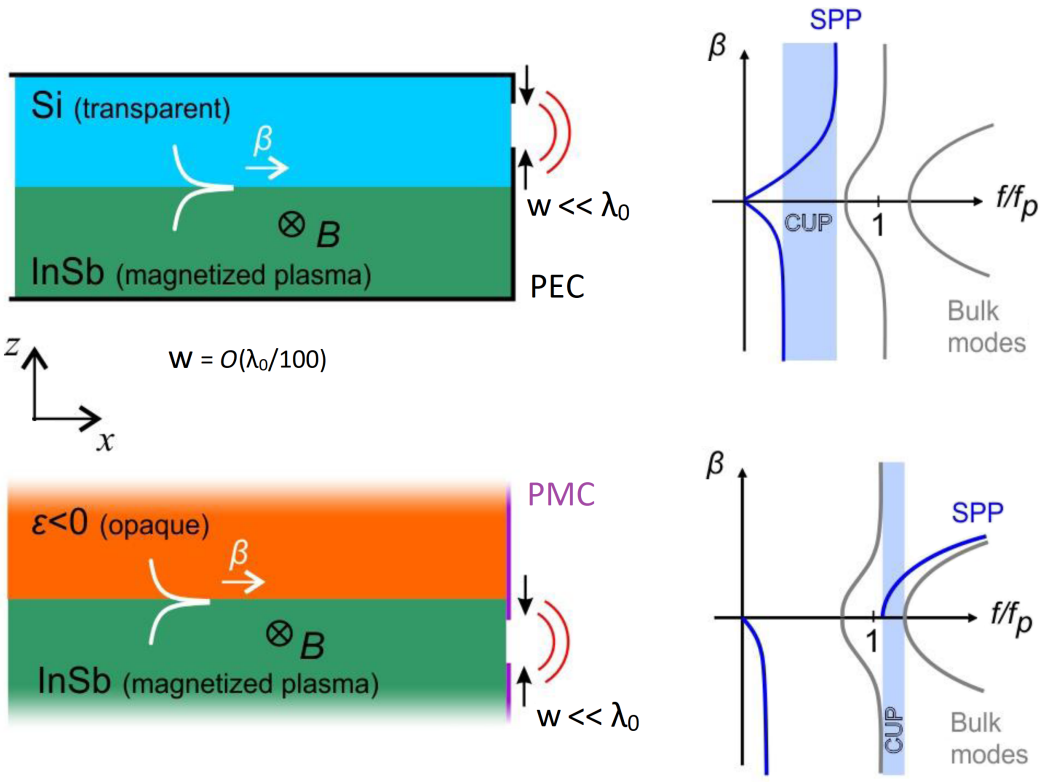


Figure 5.9: Dispersion graphs (qualitative) for the NRNT and NRTT structures. For the NRNT the $\beta = F(\omega)$ is unbounded, whereas for NRTT is not.

In Fig. 5.9 it is clearly seen the effect that the topological character causes on the dispersion of the SMPs. Concerning the NRNT structure, which is unidirectional but not topological, the propagation constant β as a function of the operational frequency is unbounded; this means that β can become large in comparison to spatial inhomogeneities and fluctuations; as a result, the SMP undergoes spatial dispersion. In contrast, in the

[17] See §1.4.3.

[18] The gap Chern number is the summation of the Chern numbers of all the modes below the bandgap.

case of NRTT structure which additionally is topological, β is bounded; thereby, as it is shown in [72] it is unaffected by nonlocal effects and does not suffer from spatial dispersion.

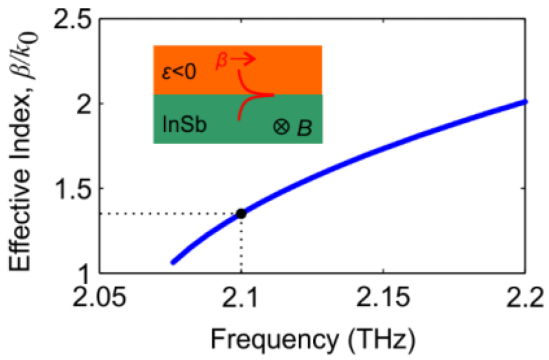


Figure 5.10: Effective refractive index of the SMPs for the NRTT structure in a real case. The applied magnetic field is -0.25 T and the CUP band Δf = 2.06 - 2.2 THz.

In Fig. 5.10 it is shown the effective refractive index, $n_{eff} = \beta/k_0$, for the NRTT structure from the simulation of a realistic case. The CUP band is

$\Delta f = 1.03\omega_p$ - $1.1\omega_p$ = 2.06 - 2.2 THz, $\omega_p = 4\pi{\cdot}10^{12}$ rad/s,

and is the only frequency region shown in the figure.
In the spotted frequency $\omega_0 = 1.05\omega_p = 2.1$ THz it is $n_{eff} = 1.3525 - i\,0.0478$.
Note that in this case the applied magnetic field is -0.25 T (has negative sign) because now the layer above InSb possesses a negative permittivity. In fact Fig. 5.10 is essentially the dispersion diagram. In this contex, it can be seen that the form of the curve is in agreement with the corresponding curve in Fig. 5.9; the same is also true for diagram in Fig. 5.7 for the NRNT structure.

5.4 Computational results in the 2D case

To test the APOTUS-HM, full-wave simulations were done using COMSOL, and some indicative results are presented here. As mentioned earlier, the permittivity of InSb is a tensor of the form (5.1); in specific, it is

$$\bar{\boldsymbol{\varepsilon}}_{\text{InSb}} = \varepsilon_0\,\varepsilon_\infty \begin{bmatrix} \varepsilon_1 & 0 & i\varepsilon_2 \\ 0 & \varepsilon_3 & 0 \\ -i\varepsilon_2 & 0 & \varepsilon_1 \end{bmatrix}, \tag{5.11}$$

where

$$\varepsilon_1 = 1 - \frac{\omega_p^2\,(\omega + i\nu)}{\omega\left[(\omega + i\nu)^2 - \omega_c^2\right]}\ , \tag{5.12a}$$

$$\varepsilon_2 = \frac{\omega_c\,\omega_p^2}{\omega\left[(\omega + i\nu)^2 - \omega_c^2\right]}\ , \tag{5.12b}$$

$$\varepsilon_3 = 1 - \frac{\omega_p^2}{\omega\,(\omega + i\nu)}\,. \tag{5.12c}$$

Unless otherwise stated, in all cases below it is set $\varepsilon_\infty = 15.6$,
plasma frequency $\omega_p = 4\pi{\cdot}10^{12}$ rad/s, cyclotron frequency $\omega_c = 0.25\omega_p$,
and loss factor $\nu = 5 \cdot 10^{-3}\omega_p$.

NRNT structure

In Fig. 5.11 it is shown the transmission coefficient as a function of the slit offset, in the NRNT structure. The operational frequency[19] is $f = 1.4$ THz.
For slit width $w = 2$ μm $= \lambda_0/100 \simeq \lambda_{eff}/20$, where λ_{eff} is the effective wavelength[20] for the excited guided mode, the maximum transmission through the slit is $T \simeq 23\%$, and is observed at a vertical offset ~ 3.75 μm.
For slit width $w = 1$ μm $= \lambda_0/200 \simeq \lambda_{eff}/40$, the transmission reaches $T \simeq 18\%$ (not shown). As is usually done in EOT studies [163, 179], this transmission is equivalent to normalized (to the incident in-the-slit-only power) transmission greater than unity, namely $N_t \simeq 1.14$, i.e., it is a truly "extraordinary" transmission [163].
The dependence of the transmission on both the slit width and offset it is presented in Fig. 5.12.

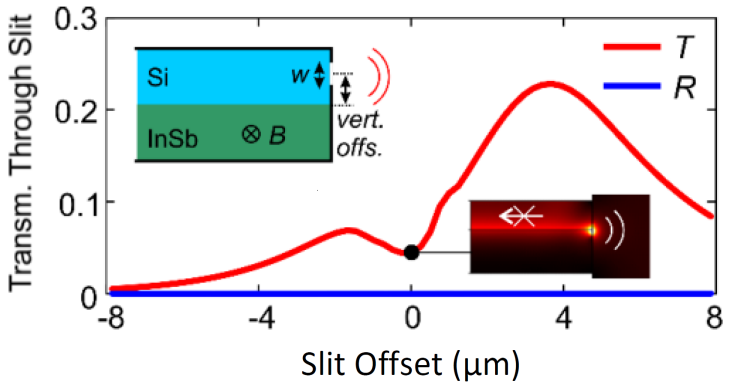


Figure 5.11: Transmission coefficient for the NRNT structure as a function of slit offset (vertical distance from the separating interface of InSb-Ag).

[19] that is, the frequency of the incident SMP.

[20] where $\lambda_{eff} = \lambda_0/n_{eff}$.

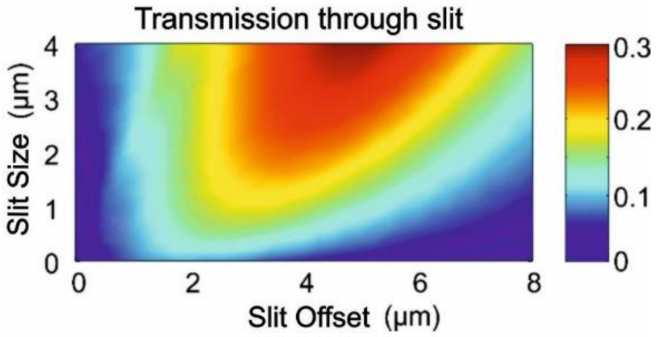


Figure 5.12: Transmission coefficient for the NRNT structure as a function of both the slit width and offset.

NRTT structure

Similarly, in Fig. 5.13 it is shown the transmission coefficient as a function of the slit offset, for the NRTT structure. The operational frequency is $f = 2.1$ THz.
For slit width $w = 1$ μm $= \lambda_0/142 \simeq \lambda_{eff}/74$, the maximum transmission through the slit is $T \simeq 25\%$, and is observed at a vertical offset ~ -0.4 μm (black curve).
For slit width $w = 2$ μm $= \lambda_0/71 \simeq \lambda_{eff}/37$, the transmission reaches $T \simeq 28.5\%$, at a vertical offset ~ 0 μm (red curve).

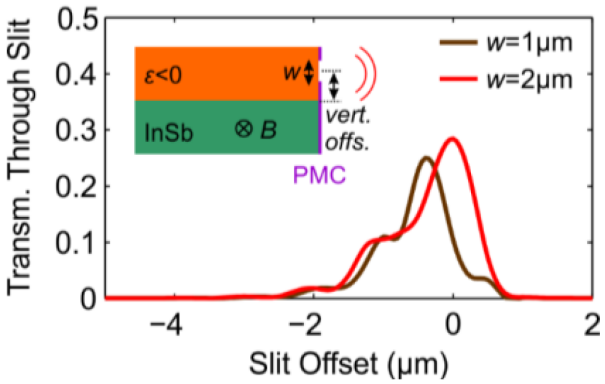


Figure 5.13: Transmission coefficient as a function of slit offset, for the NRTT structure.

In Fig. 5.14 it is presented the transmission, reflection and absorption coefficients, inside and near the CUP band, for a slit width $w = 2$ μm of the NRNT and NRTT structures. Note that in both cases the reflection coefficient inside the CUP band is virtually zero, as expected from the theory. Since the materials are not perfectly lossless[21], there is some absorption, but the transmission coefficient continues to have high values.

[21] It is reminded that the loss factor ν is not zero, see p. 128.

It can be seen that the topological structure has in general better performance : in the NRTT structure the transmission coefficient retains high values in all the CUP band, whereas in the NRNT structure decreases tending to zero.
In the examined case, for the more robust NRTT structure, the transmission through the slit reaches 33% at the frequency of 2.18 THz (cf. Fig. 5.14d); for this value the normalized (to the incident in-the-slit-only power) transmission is $N_t \simeq 1.52$, namely it is "extraordinary".

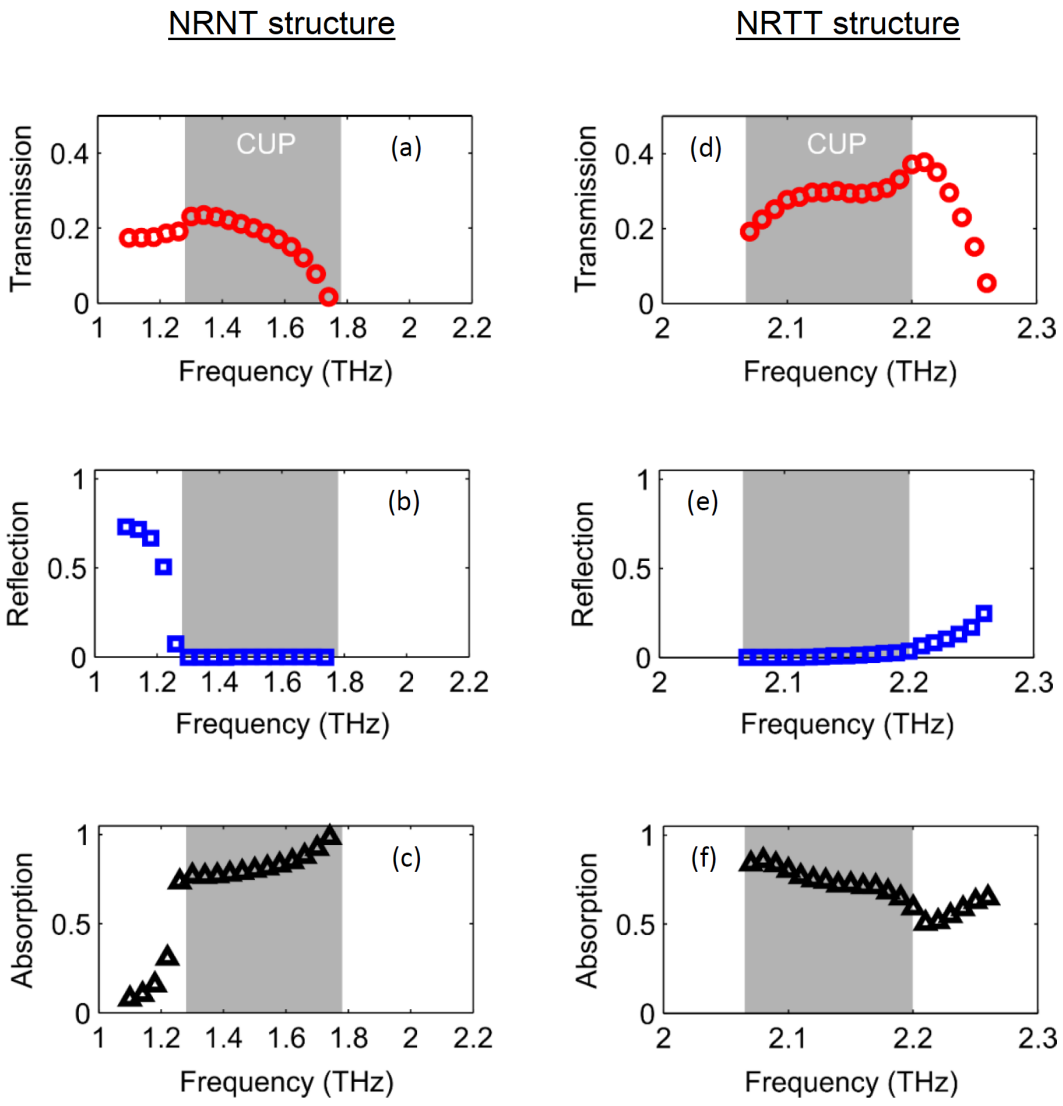


Figure 5.14: Transmission, reflection and absorption coefficients, inside and near the CUP band, for a 2 μm slit of the NRNT and NRTT structures, for the optimum respective slit positions in both cases.

Note that in all these simulations, Figs. 5.11 - 5.14, the nonlocal effects are not considered; thus, it is to be understood that robust performance as described by the theory is ultimately attained only by the NRTT structure, which is inherently topological.

It is strongly emphasized that these slit widths are in the deep subdiffractional regime (e.g., the slit dimensions are $\lambda_0/100$ or smaller) and the transmission coefficients attained are incredibly high compared to the other as far known methods.

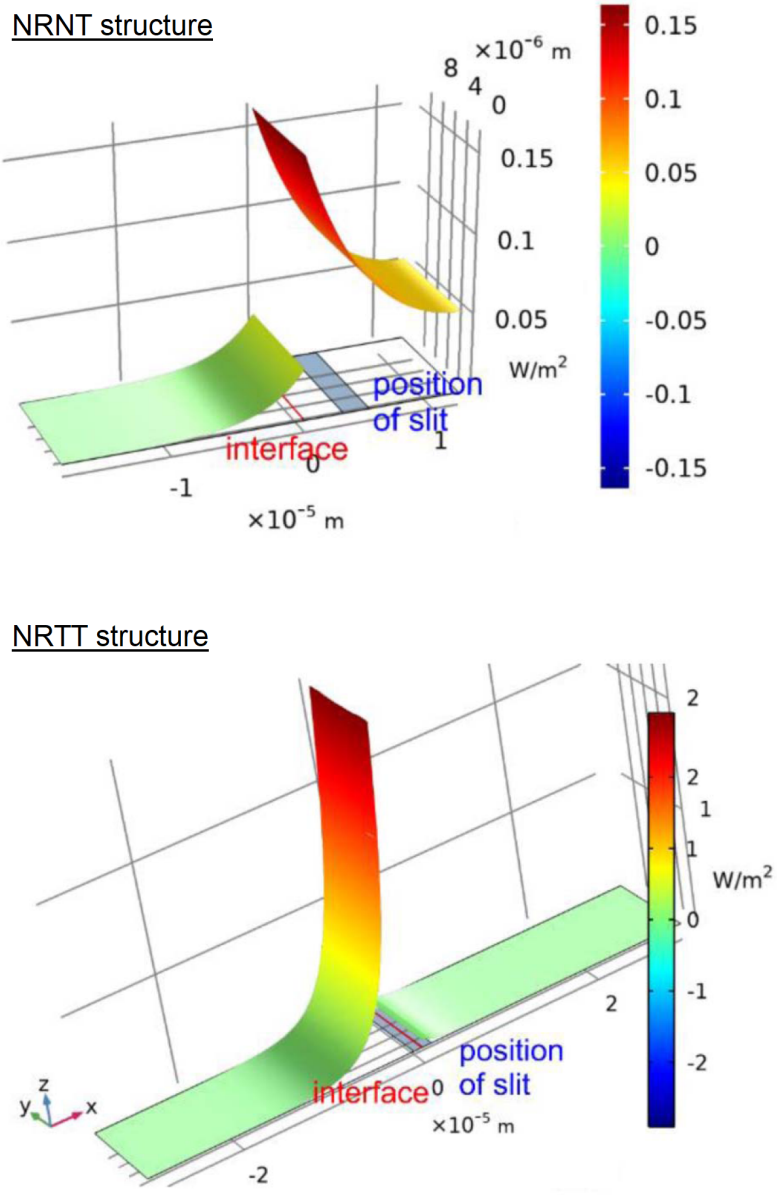


Figure 5.15: Power profile distributions for the NRNT and NRTT structures, in front of the slit. The data used are the same as in the cases of Figs. 5.11 and 5.13 respectively.

In Fig. 5.15 it is presented how the power of the guided mode is distributed to the media consisting the structure. It is reminded that the device (the waveguide) is consisted by a two-layer media. What is shown in Fig. 5.15 is how the power of the incident SMP is distributed to each layer, in front of the slit. The discontinuity in the graphs is due

to the different material on either side of the separating interface. As can be seen, for the NRNT structure the maximum power is at right side, which correpsonds to Si; for the NRTT structure is at left, that is in InSb. The importance of this diagram is that it indicates where the slit for the wave to pass should be opened. The slit should be opened to the material in which the power concentration is maximum, so that outgoing power is the maximum possible. In this case, for the NRNT structure it must be opened on the top layer, while for the NRTT at the bottom.

5.5 Computational results in the 3D case

All the previous results for APOTUS-HM concern 2D case studies; this means that the structures discussed have the geometry of a parallel-plate waveguide (extending infinitely along y-direction), and the hole is in fact an 1D slit[22]. However, the principles of APOTUS-HM presented earlier do not depend on the dimension of the structure, which means that all the crucial conclusions for realizing such a device remain untouched[23]. In this section, are presented simulation results for NRNT and NRTT structures in their 3D version, using the software COMSOL, mainly on the terahertz regime. As will be seen, all the beneficial properties of these structures continue to hold in their 3D version (and even better than their 2D counterparts !), making such devices useful for diverse applications in the broader field of Photonics. It is emphasized that achieving truly topologically unidirectional EM-based EOT in this range is challenging due to the presence of nonlocal effects [72, 73, 39]; this is a crucial issue that is thoroughly examined and addressed herein[24]. To gain a better picture on the improvement of transmission characteristics bringing by the 3D APOTUS-HM devices, their corresponding 2D versions are

[22] Although an 1D slit has a width with a deep subwavelength scale (i.e., it is very "thin"), its fundamental mode is not cut off, which means the transmission channel is not closed and EOT can take place with no need for unidirectional surface modes. However, in such a case, nothing prevents the wave to be reflected and travel backwards, away from the hole; thus, at the hole can take place backscattering, causing significant losses. This cannot happened in unidirectional structures used in APOTUS-HM. Striclty speaking, it is not fair to compare the transmission efficiency of an 1D slit with that of the Bethe's hole or NSOM probe tip, because the waveguide mode is cutoff in these typical structures whereas in an 1D slit is not; even so, the fact that the transmission through the tiny slit or hole does not depend any more on how subdiffractional they are (as long as there is transmission or tunneling through them), is a remarkable result, never before observed or attained in the entire field of EOT where the transmission through, e.g., $\sim \lambda_{eff}/50$ holes or slits is negligible, even with the use of SPPs.

[23] Note that the analytic, Temporal Coupled Mode Theory (TCMT) developed in §5.7 has no assumptions for the dimension of its formalism. Thus, TCMT, especially the conclusive Eq. (5.30), holds in the same way for 2D and 3D structures, suggesting that in such unidirectional devices the transmission through a tiny hole or slit is the maximum possible, no matter of the dimension of the model.
Note also that the extension to 3D of parallel-plate-like waveguide structures is quite straightforward and has been reported many times in the literature. For example, YIG-based 2D unidirectional waveguides, cladded with PEC metals on the upper and lower xz-planes of the guide, give rise to TE modes, where (if x is the direction of propagation) the three components of a mode are E_z, H_x and H_y. The z-direction electric field component makes it possible to cover the structure with a PEC metal on the two xy planes too, thereby making the structure 3D, with no influence at all on its one-way properties because the electric field component remains perpendicular to the PEC boundaries on the xy-planes.

[24] In contrast, in §5.4 nonlocality was not taken acnount in the simulations.

also examined and compared to them.

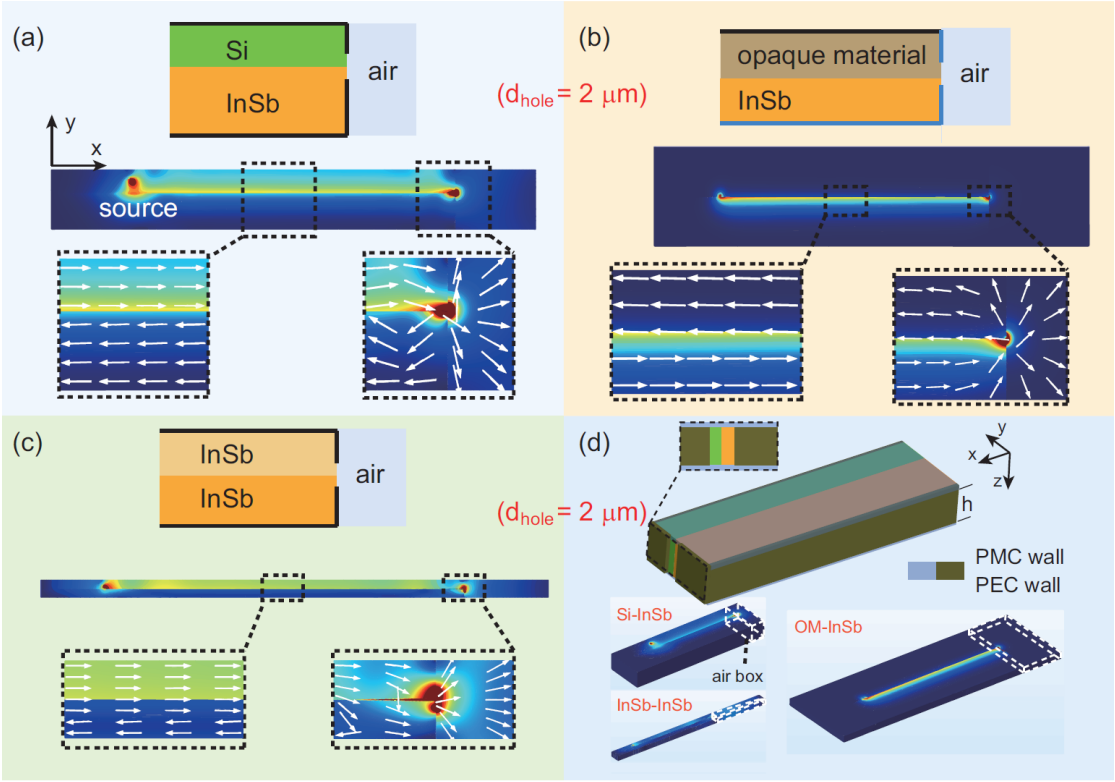


Figure 5.16: Schematic diagrams and snaptshots from simulations of three different structures with a 2 μm hole at the end.
(a) Si-InSb, (b) OM-InSb, (c) InSb-InSb.
Arrows in the snaptshots indicate the average power flow, which varies in direction based on the negative effective permittivity [247, 216].
(d) Schematic diagram and snapshots from simulations of 3D EOT, with all other parameters the same as in Fig. 5.17.

The 2D and 3D structrures to be considered are shown in Fig. 5.16. Firstly, two classic heterostructures are studied: one composed of InSb and a dielectric material (silicon here), and the other composed of InSb and an opaque material[25] (OM) with a plasma frequency larger than InSb's (e.g., $\omega_{p_m} = 2\omega_p$) [72]. As shown in Figs. 5.17a, b, both structures exhibit a unidirectional band due to the breaking of time-reversal symmetry. In particular, the Si-InSb structure exhibits a unidirectional band, characteristic of asymptotic frequencies (AFs), while the OM-InSb structure features, first, a nonreciprocal AF-type unidirectional band, and, second, a truly topological unidirectional band[26] with the surface modes having finite wavenumbers, insulating them from

[25] This must be a plasmonic material ($\varepsilon < 0$).

[26] It is possible for a material or structure to exhibit many bands, some nontopological and some truly topological. The character in each band can be different; a topological material does not necessarily exhibit topological character in all of its bands. In the case of Fig. 5.17b the structure exhibits two bands;

nonlocal effects [72].

The OM-InSb structure, in the upper boundary is surrounded by PEC, whereas in the lower boundary is surrounded by PMC; this is required for preserving the unidirectional band[27]. Indeed, the zoomed-in panel in Fig. 5.17b shows the dispersion curve if PEC is set instead of PMC in the lower boundary : then the topological unidirectional band is closed and the propagation is two-way.

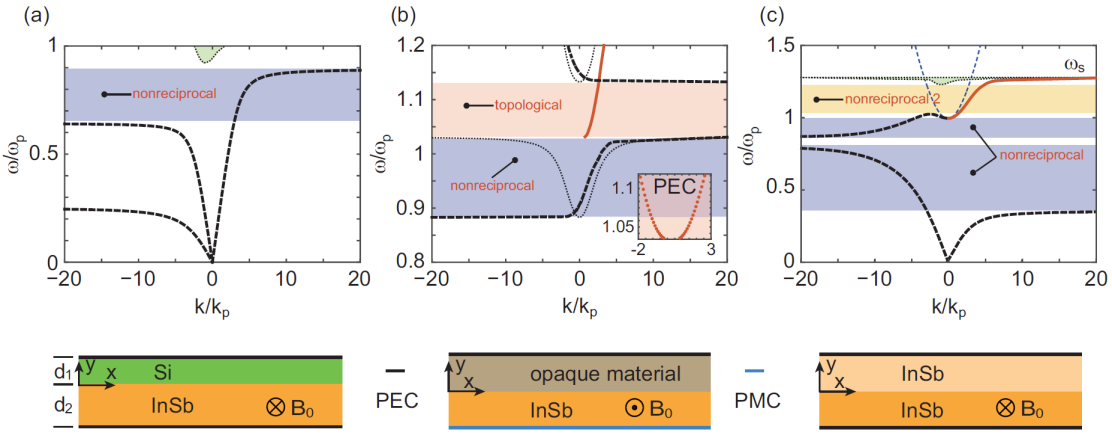


Figure 5.17: Dispersion and schematic diagrams of
(a) the Si-InSb structure with $\omega_c = 0.25\omega_p$, $d_1 = 0.08\lambda_p$ and $d_2 = 0.12\lambda_p$,
(b) the opaque material (OM)-InSb structure with $\omega_c = -0.25\omega_p$ and $d_1 = d_2 = 0.2\lambda_p$,
(c) the InSb-InSb structure with $\omega_c = 0.8\omega_p$ and $d_1 = d_2 = 0.03\lambda_p$.
The dielectric constants used are $\varepsilon_{Si} = 11.68$ and $\varepsilon_\infty = 15.6$.
Red lines indicate the topological surface modes that are immune to nonlocal effects; black dotted lines represent nonreciprocal modes propagating in only one direction in the local case.

The opaque material of the second structure is an "idealized" material, in the sense that it has the simple Drude model[28]. For the third examined structure, the opaque material has been replaced by *unmagnetized* InSb; also, both the upper and the lower cover of the structure is PEC. In Fig. 5.17c it is shown the corresponding dispersion diagram. As can be seen there are three nonreciprocal bands; only the upper one (yellowish color)

in the first, the Chern numbers are zero, hence it is simply a nonreciprocal band (nontopological); in the second, the Chern numbers are *nonzero*, hence the character of the structure is truly topological in this frequency range. In such cases the name "NRTT" (non-reciprocal, truly topological) will concern the structure when it works in the topological band.

[27] The usage of PMC in the lower boundary is necessary for the structure to exploit the topological character. If PEC is used instead of PMC, then the the topological band vanishes, as can be seen in the zoomed-in panel of Fig. 5.17b. Whether PEC or PMC must be put, is found by trials because it is difficult to find a priori which of them gives the desired result. Which of the two is appropriate has to do with the orientation of the field components.

[28] It does not matter what exactly the material is, but only that it has the simple Drude model.

is useful. This band is unidirectional, and also protected from nonlocal effects. In fact, it can be seen that just on the upper end of the band the wavenumber (red line) takes a finite number; then it becomes asymptotically infinite – but just outside of the band ! Thus, this band supports surface modes with finite wavenumber, which are therefore expected to be robust against nonlocal effects[29]. Specifically, the calculations show that the surface modes in this upper unidirectional band (yellowish color) have finite wavenumbers, with $|k| < 10k_p$, where $k_p = \omega_p/c$.

Next, the impact of nonlocal effects in the afforementioned InSb-based structures is investigated. To include the nonlocality, the hydrodynamic model is used, which arises from the presence of spatial currents **J** [72, 39, 201]. In the hydrodynamic model the currents are described by the following equation[30] :

$$\beta^2 \nabla(\nabla \cdot \mathbf{J}) + \omega\,(\omega + i\gamma)\mathbf{J} = i\omega\,(\omega_p^2\,\varepsilon_0\varepsilon_\infty \mathbf{E} + \mathbf{J} \times \omega_c \hat{z})\,, \tag{5.13}$$

where β is the nonlocal parameter, γ represents the damping rate, and $\omega_c = eB/m^*$ is the cyclotron frequency related to the external magnetic field; also, B is the intensity of the applied magnetic field, and e and m^* are the charge and the effective mass of electron respectively. In the limit of $\beta = 0$, Eq. (5.13) reduces to the classical Drude model, which implies that the current **J** at a point is determined solely by the EM field at that point. Since Maxwell's equations require that **J** has the same wavenumber as the electric and magnetic fields, the nonlocal effect is expected to have little impact on EM waves with wavenumbers satisfying the condition

$$k \ll \omega/\beta\,. \tag{5.14}$$

In this study, the value of β is set to[31] $\beta = 1.07 \cdot 10^6$ m/s for n-type InSb. For $\omega = \omega_p$, the condition (5.14) reduces to $|\bar{k}| \ll c/\beta \approx 280$, where $\bar{k} = k/k_p$. This indicates that, as discussed above, in the structure of Fig. 5.17c the SMPs of interest should be almost immune to nonlocal effects, regardless of the presence of SMPs with large k at higher frequencies (beyond the unidirectional band), near the resonance frequency ω_s of the bulk modes.

To investigate the impact of nonlocality on the propagation of SMPs in the studied structures, their properties are analysed using the software COMSOL with a nonzero nonlocal parameter ($\beta \neq 0$). For each structure of Fig. 5.17, the propagation of a pulse is examined in two cases : one with the classical Drude model, and another one with the hydrodynamic model with the aforementioned value for β. The simulations are done with the Finite Element Method. The results are shown in Fig. 5.18.

---

[29] In this case it is indifferent if the structure is topological or not; since k is finite, the surface modes are certainly protected against the nonlocal effects.

[30] Roughly speaking, in the hydrodynamic model the electrons are regarded as a fluid; equations of fluids are used to setup this model. The hydrodynamic model introduces nonlocality, which is the desired in this case.

[31] This is a typical value for β, taken from other studies in the literature involving the hydrodynamic model, e.g. [72].

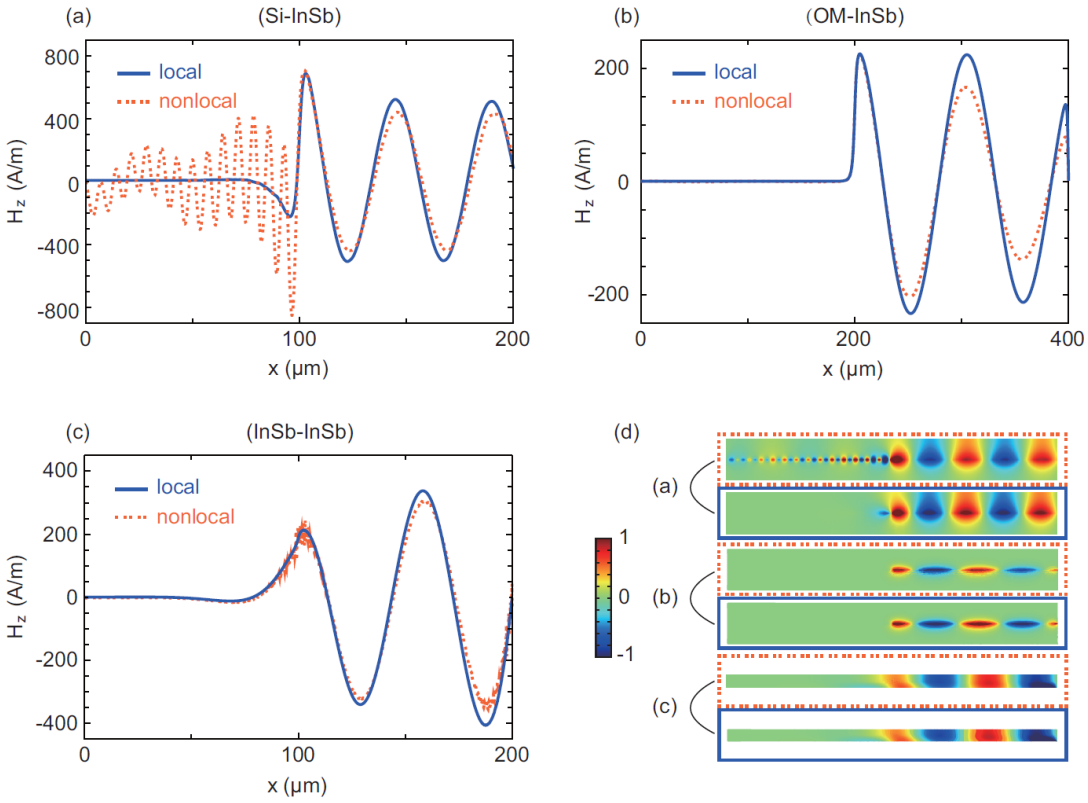


Figure 5.18: The magnetic field distribution obtained from FEM simulations of three different structures :
(a) Si-InSb, (b) OM-InSb, and (c) InSb-InSb.
The simulations were performed at three different frequencies : $0.8f_p$, $1.05f_p$, and $1.1f_p$, where f_p is the plasma frequency.
The damping rate γ was set to $\gamma = 5 \cdot 10^{-3}\omega_p$, and the nonlocal parameter β was set to $\beta = 1.07 \cdot 10^6$ m/s.
(d) Snapshots of the magnetic field distribution in the three structures under local conditions (inside the blue rectangle) and nonlocal conditions (inside the red rectangle).

Fig. 5.18 shows that, as expected, in the first Si-InSb structure the magnetic field distribution exhibits significant spatial dispersion, resulting in backward modes within the nonreciprocal unidirectional band limited by AFs. In the remaining structures, the SMPs maintain their unidirectional character, and the magnetic fields exhibit negligible differences between the local and nonlocal cases. As a result, it is concluded that the Si-InSb structure is susceptible to realistic nonlocal effects because of the presence of large wavenumber k in the unidirectional band (see Fig. 5.17a).
In contrast, the OM-InSb and InSb-InSb structures sustain robust SMPs, as they have relatively small k values ($k < 10$) within their respective unidirectional bands, making them essentially immune to nonlocality.

Next, the transmission efficiency of the 2D and 3D structures of Fig. 5.16 is studied. Two cases are examined: in the first the diameter of hole is $d_{hole} = 2$ μm $\simeq \lambda_0/100$, whereas in the second it is $d_{hole} = 1$ μm $\simeq \lambda_0/200$. The results are shown in Figs. 5.19 and 5.20.

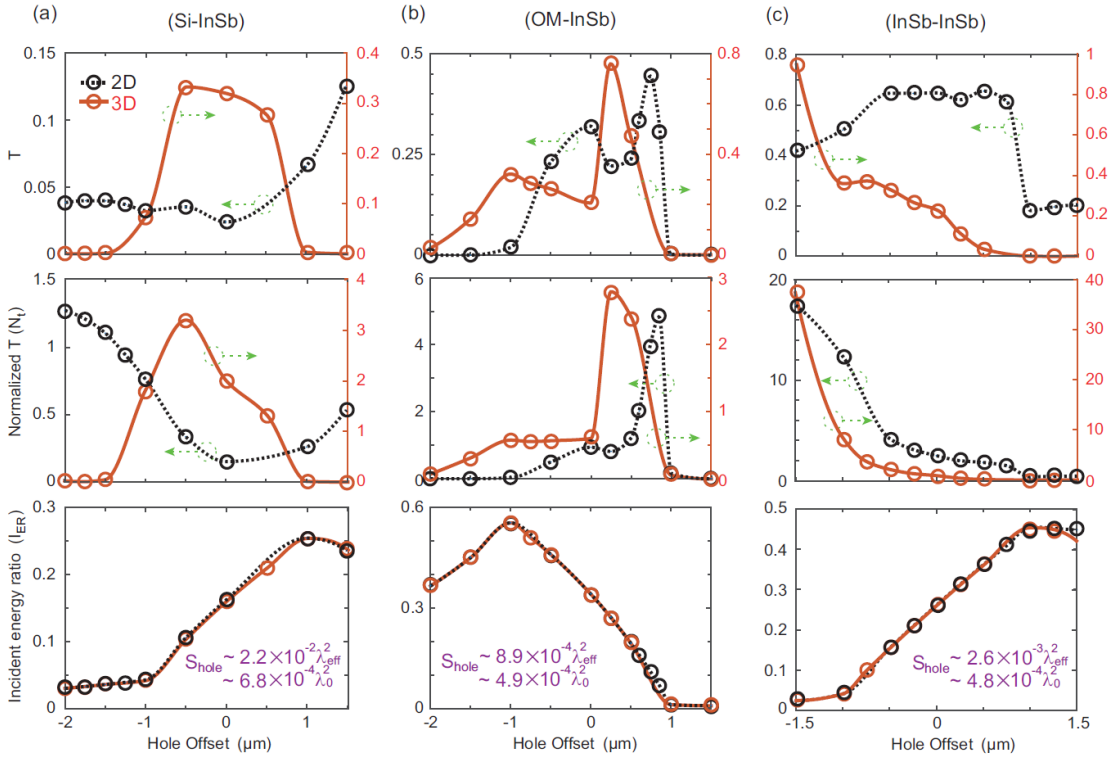


Figure 5.19: The transmission coefficient T, the *normalized* transmission coefficient N_t, and the incident energy ratio I_{ER}, as functions of the hole offset, for the three structures of Fig. 5.16.
The hole has a diameter $d_{hole} = 2$ μm $\simeq \lambda_0/100$, where λ_0 is the vacuum wavelength.
Note that the climax for the 2D and the 3D version of the structures are different (left and right sides of the diagrams.)

In Fig. 5.19 it is shown the transmission coefficient T, the normalized transmission coefficient N_t, and the incident energy ratio[32] I_{ER}, plotted as functions of the hole-offset, for the three structures of Fig. 5.16, in both the 2D and 3D versions.

Concerning the 2D versions, in agreement with a previous study [14], the Si-InSb structure exhibits a quite low transmission efficiency[33] ($T^{max} < 12\%$; see upper panels in Fig. 5.19). In contrast, the OM-InSb and InSb-InSb structures exhibits significantly higher transmission efficiencies, with $T^{max} \approx 45\%$ and $T^{max} \approx 65\%$, respectively. This difference in transmission efficiency can be attributed to the effective refractive

[32] See Eq. (5.15) in p. 138.

[33] Of course it is very high in comparison with the conventional methods.

indices[34] of the structures, which are $n_{eff} \approx 5.7$ (with $f = 0.8f_p$ and $k \approx 4.56k_p$), $n_{eff} \approx 1.35$ (with $f = 1.05f_p$ and $k \approx 1.42k_p$), and $n_{eff} \approx 2.31$ (with $f = 1.1f_p$ and $k \approx 2.54k_p$) for the Si-InSb, OM-InSb, and InSb-InSb structures, respectively.

Indeed, physically, a wave with a large effective refractive index, such as the one analyzed in Fig. 5.16a, cannot efficiently coupled with the evanescent wave in the air ($n = 1$) surrounding the hole, due to the larger refractive index difference.
To examine the presence or not of EOT, the normalized transmission coefficient[35] N_t is also calculated; it was found (see Fig. 5.19, middle panels) $N_t^{max} \approx 1.25$ (Si-InSb), $N_t^{max} \approx 5$ (OM-InSb), and $N_t^{max} \approx 18$ (InSb-InSb).

Next, it is analyzed the topological or nonreciprocal EOT in the 3D structures. In the InSb-based unidirectional waveguide, the guiding modes are transverse-magnetic (TM) modes. Therefore, the presence of lateral PMC walls [9] in the xy-plane do not destroy the topological character, nor the EOT phenomenon[36].
The configuration of the examined 3D structures and the position of the hole is shown in Fig. 5.16d and Fig. 5.17 (regarded as 3D in this case). The thicknesses of the 3D structures are $h = 12$ μm $\simeq 0.064\lambda_0 \simeq 0.36\lambda_{eff}$, $h = 5$ μm $\simeq 0.035\lambda_0 \simeq 0.047\lambda_{eff}$, and $h = 4.5$ μm $\simeq 0.033\lambda_0 \simeq 0.015\lambda_{eff}$, respectively. Finite element simulations were done on these deep subwavelength structures, and verified the preservance of the unidirectional propagation.
For a hole with $d_{hole} = 2$ μm, and for no hole offset, it was found 3D EOT with $N_t \approx 2$ in the Si-InSb structure under the present conditions (see Fig. 5.19a, middle panel).

Furthermore, it is crucial to also study how the waveguide parameters, particularly the hole offset, affect the 2D and 3D topological/nonreciprocal EOT, as optimizing these parameters can lead to dramatically improved throughput performance[37]. To investigate this, it was performed a series of simulations on InSb-based structures with different hole offsets, in both the 2D and 3D configurations. Here, it is useful to define the incident energy ratio I_{ER} to characterize the energy distribution in the present magneto-optical structures. Specifically, I_{ER} is defined as the ratio of the incident energy over the hole region E_{hole} to the total incident energy E_{total}, that is

$$I_{ER} = \frac{E_{hole}}{E_{total}} \,. \tag{5.15}$$

This parameter is used to emphasize the significance of the energy distribution in the structures. The value of N_t can then also be calculated using the following equation :

$$N_t = \frac{T}{I_{ER}} \,. \tag{5.16}$$

It is evident from (5.16) that EOT can occur in cases with low I_{ER} and/or high T. Ideally, it is required relatively high T and low I_{ER}, which would make these structures excellent candidates for building sensitive optical devices, including NSOM operation –

[34] It is reminded that $n_{eff} = \beta/k_0$ and $\lambda_{eff} = \lambda_0/n_{eff}$.

[35] Defined as the ratio of transmitted power to the power that incidents *only* on the slit or hole.

[36] Because of the orientation of the field components, the presence of PMC does not affect the field (there is a normal field component which remains unaffected from PMC).

[37] Remember Fig. 5.12 in p. 129.

only that here, being inherently unidirectional, the devices do not require elongated tapering, thereby avoiding unnecessary propagation losses, and leading to much enhanced throughput.

As already mentioned, for the structures with $d_{hole} = 2$ μm $\simeq \lambda_0/100$, the results are presented in Fig. 5.19. For the 2D Si-InSb structure, the calculations shown in Fig. 5.19a reveal that most of the energy is concentrated in the Si layer, as the I_{ER} at a positive hole offset (> 0 μm) is typically larger than that in the region of negative hole offset (< 0 μm). However, due to a lack of momentum matching, only a small part of the energy can successfully escape to the air. When the hole is located in the lower InSb layer, N_t can become greater than 1, with a maximum of 1.25, but for $T < 5\%$. Interestingly, for no hole offset, the 3D structure actually outperforms its 2D counterpart. Specifically, as shown in Fig. 5.19a, it achieves $N_t > 3$ with $T > 30\%$.

For the 2D OM-InSb structure, Fig. 5.19b shows that the energy in the opaque material layer is lower than that in the InSb layer. Unlike the Si-InSb case, 2D EOT with $N_t \approx 5$ can occur at relatively high T ($T \approx 45\%$), while in the 3D case it is $N_t \approx 3$ ($T \approx 75\%$).

Remarkably, for the third structure (InSb-InSb), Fig. 5.19c shows extremely high EOT in both the 2D and 3D cases. For the 2D structure the transmission coefficient is $T^{max} \approx 65\%$ and for the 3D one $T^{max} \approx 95\%$; the normalized values are $N_t^{max} \approx 20$ and $N_t^{max} \approx 40$, respectively. The calculation of I_{ER} indicates that the energy distribution is similar to that in Fig. 5.19a, but here the attained transmission is substantially higher.

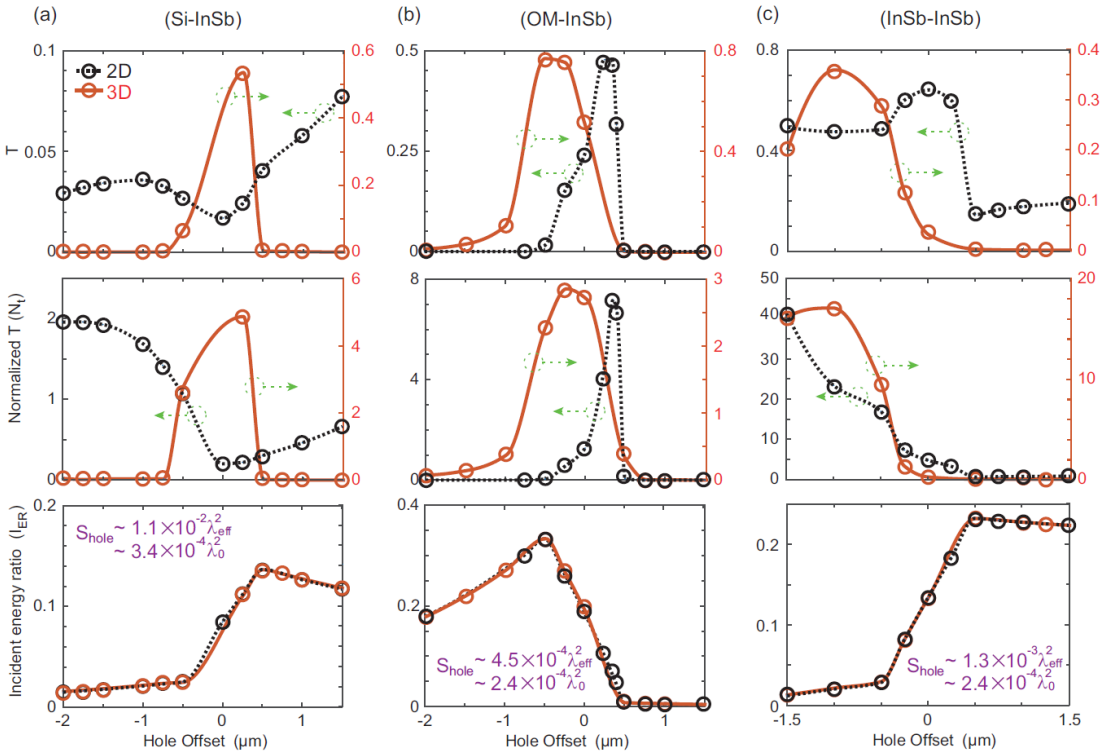


Figure 5.20: The same as Fig. 5.19 but here the diameter of the hole is $d_{hole} = 1$ μm $\simeq \lambda_0/200$, where λ_0 is the vacuum wavelength.

Another parameter that is expected to affect the transmission is the diameter of the hole. As mentioned earlier, the case for a hole with smaller diameter was also examined for all the 2D and 3D structures. Specifically, all the study of Fig. 5.19 was repeated, setting the diameter of the hole to $d_{hole} = 1$ μm $\simeq \lambda_0/200$. The results are shown in Fig. 5.20. For the Si-InSb structure, it is observed that when $d_{hole} = 1$ μm, waves cannot easily out-coupled to air in the 2D structure, resulting in a $T^{max} < 10\%$. However, for the 3D version, it was found that $T^{max} > 50\%$. Also it is observed that the energy ratios I_{ER} are almost the same between the 2D and 3D structures. Thus, due to the relatively larger T in the 3D case, higher values of N_t are achieved in the 3D structure compared to the 2D one, with the N_t^{max} reaching 4.88 for a hole offset ≈ 0.25 μm. Note from Fig. 5.20a that the 3D Si-InSb structure typically exhibits very low T or N_t for hole offsets different than the central (= 0 μm) value; thus, in this case it is performance-wise to have a hole with zero offset.

For the OM-InSb structure, as shown in Fig. 5.20b, the transmission efficiencies in both the 2D and 3D configurations are not strongly affected by the change in d_{hole} . The maximum values of T and N_t are nearly the same as with the previous ($d_{hole} = 2$ μm, Fig. 5.19b) case. Specifically, for the 2D case it is $N_t^{max} \approx 7$ and $T^{max} \approx 50\%$; for the 3D case it is $N_t^{max} \approx 3$ and $T^{max} \approx 75\%$.

Finally, for the InSb-InSb structure, it is found that the maximum N_t reaches very large values : for the 2D structure it is $N_t^{max} \approx 41$ (for a $T^{max} \approx 49\%$), while for the 3D structure it is $N_t^{max} \approx 17$ (for a $T^{max} \approx 36\%$). In other words, the robust nonreciprocal nature of this device (immune to nonlocal effects, as discussed above) and the absence of back-reflections, lead to excellent 3D EOT performance (through a single hole), even in the deep subwavelength scale, where $d_{hole} \ll \lambda_0$. Also, note that for the Si/OM-InSb structures the maximum N_t is always achieved when the hole offset is anywhere in the range $(-d/2, d/2)$, but in the InSb-InSb structure the maximum N_t is attained when the hole is located entirely within the lower layer. This is because, in this case, the incident wave can efficiently couple into the air while I_{ER} remains small.

the transmission coefficient in the unidirectional band

In the above studies (Figs. 5.19 and 5.20) the transmission coefficient was calculated for a specific operational frequency, within the unidirectional propagation band. It is reminded here that the transmission coefficient is given by the formula

$$T = \frac{2\gamma_T}{2\gamma_T + \gamma_0}\,, \tag{5.9}$$

extracted with the Temporal Coupled Mode Theory, see §5.7 and (5.30). The physical meaning of (5.9) has been discussed thoroughly in §5.3.1. Eq. (5.9) seems very simple, but this is somewhat deceivable. As T depends only on γ_0 and γ_T, at first sight it would be expected that T is constant within the unidirectional band, as shown in Fig. 5.5. However, in reality this is not the case; γ_0 is the decay rate of the wave due to dissipative losses, and depends on the frequency of the wave propagating in the medium; these losses are imposed by the model of medium (Drude model or whichever) and are

a function of frequency[38]. Furthermore, the spatial profile of the field also varies with frequency; this affects indirectly the tunneling rate γ_T of the wave. Thereby, γ_0 and γ_T, and thus T, vary with frequency in the unidirectional band; therefore, in contrast to what naively expected, T is not constant in the unidirectional band.

Also, note that the geometry of the hole (and in specific the hole offset) is indirectly but definitely incorporated in (5.9) through γ_T : the more "convenient" the geometry is for tunneling, the larger the γ_T (and thus the T) is. This is the reason for optimizing the position of the hole offset in the configurations above.

These considerations were tested and confirmed for the three herein structures. In Fig. 5.21 is shown the normalized transmission coefficient N_t as a function of frequency, in the unidirectional band. In all cases, N_t is quite high, having values that agree in order with the values observed in Fig. 5.19 and 5.20, but it is not constant. As shown, in the Si-InSb and OM-InSb structures, 3D EOT is achieved within a portion of the respective unidirectional bands; in the InSb-InSb structure (which has the most practicality) 3D EOT is achieved throughout the entire unidirectional band.

The influence of the hole offset was also investigated; as can be seen in case of InSb-InSb structure, its role can be crucial for achieving intense EOT and high values of T.

These results also verify that although T and N_t may have significant fluctuations, they retain the broadband behavior; that is, the structures remain broadband in a large portion (or even in all) the unidirectional band.

In conclusive, clear 3D EOT was observed in these structures, at different frequencies, even when the hole position was kept unchanged for all cases. This represents a stringent condition given the complex relationship between the operating frequency, hole offset, hole size, and the maximum normalized transmission. Thus, by appropriate choice of the structural (device) parameters, it is possible to achieve 3D EOT throughout an entire (broadband) unidirectional band.

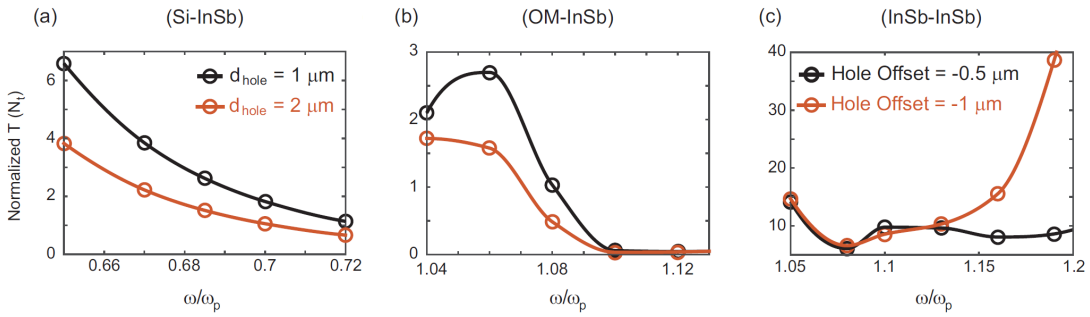


Figure 5.21: Normalized transmission coefficient as a function of operating frequency in the corresponding unidirectional band. The configuration of the structures is 3D.

[38] except if the medium has a constant ε, like Si.

5.6 Summary of the 3D results and general conclusions

Summarizing, it was introduced and then studied in some detail three different types of 3D structures capable of attaining robust (nonreciprocal or topological) EOT through single holes : PEC-Si-InSb-PEC (nonreciprocal), PEC-OM-InSb-PMC (topological), and PEC-InSb-InSb-PEC (robustly nonreciprocal, immune to nonlocal effects). Using finite element simulations, it was confirmed the existence of nonreciprocal/topological (one-way) waves in the proposed structures. The unidirectional surface waves sustained by the InSb-InSb device have relatively small wavenumbers ($k < 10k_p$), making them essentially immune to nonlocal effects. Because of the absence of back-reflections and the avoidance of the need for elongated tapering, the proposed structures exhibit nonreciprocal/topological 3D EOT with high absolute (T) and normalized (N_t) transmissions, even in the deep subdiffractional regime, and in a large portion or even in the entire unidirectional band (that is being broadband).

Specifically, for a hole surface of just $S_{hole} \approx 4.8 \cdot 10^{-4}\lambda_0^2$ ($\approx 2.6 \cdot 10^{-3}\lambda_{eff}^2$), it was found that the 3D InSb-InSb structure allows for near-perfect absolute transmission through the hole ($T \approx 95\%$) and for normalized T far exceeding unity ($N_t \approx 40$, EOT regime; see Fig. 5.19c).

The main advantages of the APOTUS-HM can be summarized as following :

- Very high transmission coefficient in comparison to other techniques, especially in the deep subdiffractional regime; e.g., in holes $d \leq \lambda_0/100$:
  gold-coated tapered optical fibers used in NSOM : $T \leq \sim 10^{-4}\,\%$,
  APOTUS-HM : $T \geq \sim 30\,\%$,
- Does not require perfect lossless material to achieve high transmission,
- Permits 3D isotropic effective refractive indices,
- It is immune to spatial dispersion, structural imperfections/inhomogeneities, and surface roughness (the propagation is robust and stable),
- Its implementation is easier in comparison to other techniques.

Although lacking yet the experience of construction and testing a real APOTUS-HM device, the limitations of the method as seems from the theoretical study and the simulations are very little and non-critical. In specific, the main difficulties are the following :

- Practically, the transmission coefficient is less than 100% due to losses in the material and to the impedance matching at the exit,
- The maximum output power is limited by heating in the neighborhood of the hole,
- High-precision numerical study is difficult due to the “exploding” of the field in the very close neighborhood of the hole.

Of the above, the potential heating in the neighborhood of the hole is the most serious problem; if the tunneling time is large (small γ_T) and the power of the wave to be focused high, the device could be burnt. For this reason, the possibility of a cooling mechanism could be examined.

The exaggeration of the field very close to the hole poses a significant difficulty to the simulation of tunneling. Just in front of the hole, before tunneling, the magnitude of the field increases abruptly two to three orders (or even more), see Fig.5.22; this demands a very dense computational mesh to catch correctly the evolution and tunnelling of the wave, in most cases in the limit of the available computational resources.

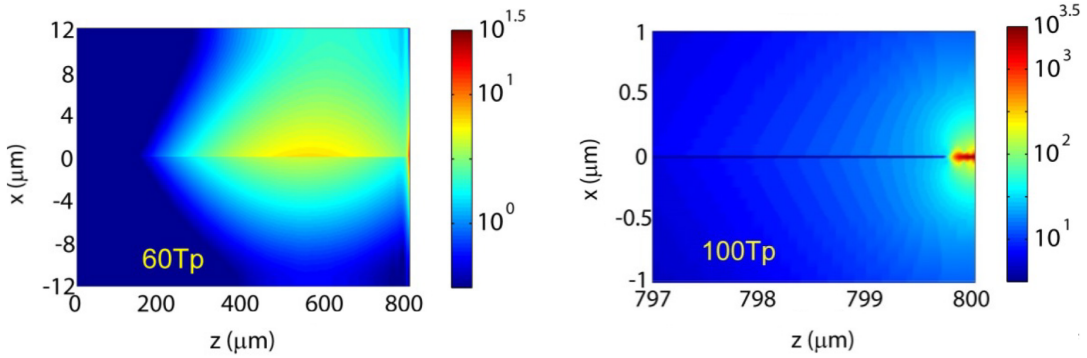


Figure 5.22: The problem of "exploding field" in front of the hole.

However, these difficulties are not limitations in principle and can be overcome more or less easily. Thus, the proposed structures may form an entirely new platform for high performance micro-/nanoscopy, heat-assisted magnetic recording, sensing, detection, enhancement of spontaneous emission and photoluminescence, and integrated photonic communication applications where there is a dire need for efficiently bringing light and electromagnetic waves to deep subwavelength scales.

5.7 Temporal Coupled Mode Theory and transmission coefficient

Temporal Coupled Mode Theory (TCMT) is an approximation technique concerning the coupling of modes in resonant systems. It is essentially a perturbation analysis and, since the resonance phenomenon can take many forms, TCMT can be used in many physical systems where resonance is involved. Among others, it can be used to treat optical nonlinearities, the interaction between optical and acoustic waves, etc. In the case of APOTUS-HM, it is used to obtain the transmission coefficient of the NRNT and NRTT structures. TCMT will not be presented here but only some fragments just necessary to obtain the transmission coefficient. A quite detailed presentation of TCMT for some systems is discussed in [96].

The NRNT and NRTT structures can be considered as cavities with in-/out-coupling channels, where power waves pass and resonate inside the cavity. The phenomenon inside them is regarded as the incidence of a waveguide wave on a resonator and time evolution of the amplitude of an excitation in the resonator [96]; in this regime the TCMT can be applied. In general, such a system (optical or not) is constrained by energy conservation, time-reversal symmetry, and reciprocity – and these are the principles that are used in TCMT.

Let be considered a cavity of n ports (in-/out-coupling channels), where k_m and d_m are the in- and out-coupling coefficients respectively, (m being an integer $\in [0, n]$); also, let $|s_{m+}|^2$, $|s_{m-}|^2$ be the in-/out-going wave powers; ω_0 the resonance frequency, and γ the total decay rate of the excited mode of amplitude a, with $|a|^2$ beeing the energy of that mode[39]. It can be prooved [96] that the evolution of the excited mode in the cavity is governed by the equation

$$\frac{da}{dt} = (i\omega_0 - \gamma)a + \mathbf{k}^{\mathrm{T}}\,\mathbf{s}_+\,, \tag{5.17}$$

where $\mathbf{k}$ and $\mathbf{s}_+$ are column vectors with the k_m and s_{m+} factors mentioned above.
The out-coupled waves are governed by the relation

$$\mathbf{s}_- = \mathbf{C}\,\mathbf{s}_+ + \mathbf{d}a\,, \tag{5.18}$$

where $\mathbf{C}$ is the matrix for direct reflection process (pathway), and $\mathbf{d}$ the vector with the aforementioned d_m coefficients.
From the conservation of energy[40] in this system it is [96, 221, 281]

$$\frac{d|a|^2}{dt} = |\mathbf{s}_+|^2 - |\mathbf{s}_-|^2\,. \tag{5.19}$$

Without harming the generality it can be set $\mathbf{s}_+ = \mathbf{0}$;

[39] All these are notations usually adopted in TCMT.

[40] Or using Poynting's theorem.

then, substituting (5.17) to (5.19) gives[41]

$$(-i\omega_0 - \gamma)a^* a + a^*(i\omega_0 - \gamma)a = -a^* \mathbf{d}^* \mathbf{d} a \,, \tag{5.20}$$

which, since $a \neq 0$, immediately results in

$$\mathbf{d}^* \mathbf{d} = 2\gamma \,. \tag{5.21}$$

In this stage, the time evolution process described by (5.17) is reversed;
that is, the following replacements are done : $a \to \tilde{a}$, $\mathbf{s}_+ \to \tilde{\mathbf{s}}_+$, $-\gamma \to \gamma$, $\mathbf{k}^{\mathrm{T}} \to \tilde{\mathbf{k}}^{\mathrm{T}}$.
For an incident wave $\tilde{\mathbf{s}}_+ = \mathbf{d}^* \tilde{a}$ of (positive) frequency $i\omega_0$ in-coupled to the mode a it holds

$$\frac{d\tilde{a}}{dt} = (i\omega_0 + \gamma)\tilde{a} \,. \tag{5.22}$$

Inserting this in (5.17) gives

$$(i\omega_0 + \gamma)\tilde{a} = (i\omega_0 - \gamma)\tilde{a} + \tilde{\mathbf{k}}^{\mathrm{T}} \mathbf{d}^* \tilde{a} \,, \tag{5.23}$$

which immediately results in[42]

$$\tilde{\mathbf{k}}^{\mathrm{T}} \mathbf{d}^* = 2\gamma \quad \text{or} \quad (\tilde{\mathbf{k}}^{\mathrm{T}})^* \, \mathbf{d} = 2\gamma \,. \tag{5.24}$$

Comparing (5.21) and (5.24) it is

$$\tilde{\mathbf{k}} = \mathbf{d} \,. \tag{5.25}$$

Concerning the NRNT and NRTT structures, these are two-port systems. Assume now, that the incident SMP pulse carries power $|s_0|^2$; then, the amplitude a_q of an excited q mode within the complete unidirectional propagation (CUP) band, localized at the terminating interface, will vary with time as

$$\frac{da_q}{dt} = i\omega_q a_q - (\gamma_R + \gamma_T + \gamma_0)a_q + \kappa s_0 \,, \tag{5.26}$$

where

γ_R is the decay rate of the mode in the backwards direction (where the SMP is coming from),

γ_T is the tunneling rate of the mode through the slit (in the forward direction),

γ_0 is the decay rate of the mode because of dissipative losses,

κ is the in-coupling coefficient.

---

[41] a^* means the complex conjugate of a, and similarly for the rest symbols.

[42] It is reminded that γ is real.

In other words, in (5.17) it is set $\gamma = \gamma_R + \gamma_T + \gamma_0$,
and $\mathbf{k}^{\mathrm{T}}$ to κ since this is a two-port system.

Setting $\gamma_R = 0$ (i.e., no back-reflections) as is the case in broken time-reversal symmetry, and $\mathbf{d} = (d_i = 0, d_j)$, where d_i is the (zero) decay of the mode in the backwards direction, then, for a lossless ($\gamma_0 = 0$) hot spot at the end (5.24) and (5.25) give[43]

$$d_j = \sqrt{2\gamma_T} = \kappa\,. \tag{5.27}$$

Using (5.27) and the above assumptions in (5.18), for the transmitted through the slit field (in free space), eventually it is obtained

$$s_T = t_D\, s_0 + \sqrt{2\gamma_T}\, a_q\, e^{i\phi}, \tag{5.28}$$

where t_D is the direct (small, for deep subwavelength slits) transmission coefficient of the SMP through the slit, and ϕ the phase difference between the in-coupling and out-coupling processes of the resonant mode. As a result, in general, the power transmission through the slit contributed by this mode at $\omega = \omega_q$ is

$$T_{a_q} = \frac{|s_T|^2}{|s_0|^2} = \left| t_D + \frac{\sqrt{2\gamma_T}\,\kappa e^{i\phi}}{\gamma_R + \gamma_T + \gamma_0} \right|^2 , \tag{5.29}$$

where t_D is the direct transmission coefficient of the SMP through the slit.
After some manipulations, the contribution of the a_q mode[44] to the transmission through the deep subdiffractional slit is eventually obtained :

$$T_{a_q} \simeq \frac{2\gamma_T}{2\gamma_T + \gamma_0}\,. \tag{5.30}$$

In deriving the above relation, it was assumed that the direct transmission coefficient t_D through the deep subdiffraction slit is negligible (as is the case), that is $t_D \simeq 0$.

[43] It is reminded that d_j and κ are complex.

[44] at $\omega = \omega_q \in \Delta\omega$, where $\Delta\omega$ is the CUP band.

5.8 Perfect Magnetic Conductor (PMC) realization

5.8.1 Introductory remarks on PMCs and likewise materials

An issue that passed unnoticed in the above discussion of APOTUS-HM, but of crucial importance for the construction of the NRTT structure in reality, is the realization of the perfect magnetic conductor (PMC). It is reminded that in the NRNT structure the termination wall with the tunneling hole is a PEC, whereas in the case of NRTT is a PMC (see Fig. 5.9). The use of PEC or PMC is imposed from the configuration exhibiting the SMP wave in the two structures respectively[45]. Any good metal can be considered approximately as PEC, but the case of PMC is not trivial as PMCs do not exist in nature at all. In this section it will be discussed very briefly the realization and properties of PMCs.

Like an electric conductor, a magnetic conductor is a material that when it is subjected to an EM field, in its interrior both the electric and magnetic fields vanish. Concerning the boundary conditions, it is reminded that on the surface between two different materials it holds in general :

$$-\hat{\mathbf{n}} \times (\mathbf{E}_2 - \mathbf{E}_1) = \mathbf{M}_S \,, \tag{5.31a}$$

$$\hat{\mathbf{n}} \times (\mathbf{H}_2 - \mathbf{H}_1) = \mathbf{J}_S \,, \tag{5.31b}$$

$$\hat{\mathbf{n}} \cdot (\mathbf{D}_2 - \mathbf{D}_1) = q_{eS} \,, \tag{5.31c}$$

$$\hat{\mathbf{n}} \cdot (\mathbf{B}_2 - \mathbf{B}_1) = q_{mS} \,, \tag{5.31d}$$

where q_{eS} and q_{mS} is respectively the surface electric and magnetic charge, and the rest of the notation is shown in Fig. 5.23.

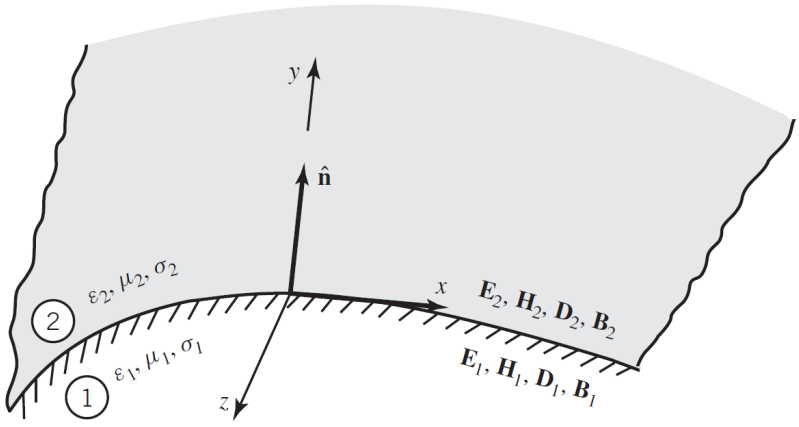


Figure 5.23: Notation used in the general boundary conditions for EM field on the surface of two materials.

[45] Note that in NRNT the upper layer (Sb) has $\varepsilon > 0$, whereas in NRTT the upper layer is plasmonic and has $\varepsilon < 0$.

Let consider the medium 1 to be a PMC; then it holds $\mathbf{E}_1 = \mathbf{H}_1 = \mathbf{0}$, $\mathbf{J}_S = \mathbf{0}$, $q_{eS} = 0$, and the relations (5.31) become

$$-\hat{\mathbf{n}} \times \mathbf{E}_2 = \mathbf{M}_S \,, \tag{5.32a}$$

$$\hat{\mathbf{n}} \times \mathbf{H}_2 = \mathbf{0} \,, \tag{5.32b}$$

$$\hat{\mathbf{n}} \cdot \mathbf{D}_2 = 0 \,, \tag{5.32c}$$

$$\hat{\mathbf{n}} \cdot \mathbf{B}_2 = q_{mS} \,. \tag{5.32d}$$

Eq. (5.32b) means that in a PMC the *tangential magnetic* components of the field vanish next to the surface; Eq. (5.32c) means that the *normal electric field* component vanishes too[46]. Also, Eqs. (5.32a, d) mean that the magnetic charge goes to the surface of the PMC, thus creating a magnetic current density that resides on a very thin layer at the surface.

PMCs do not exist in nature; however, they are often used hypothetically to create electromagnetic systems equivalent to the originals (i.e., they give the same results) and are more easy to handle. Materials behaving approximately as PMCs over a limited frequency band can be constructed artificially. PMCs belong to a wide family of structures known as Photonic Band-Gap (PBG) structures; these are periodic structures (in one, two or three dimensions), dielectric and conductive, which have the ability to control the electromagnetic radiation so as to prevent propagation in specific frequency bands (band gaps). PBG designation originally referred to structures for applications in Optics (and at optical frequencies) [271, 37], but gradually was greatly expanded and now includes Electromagnetic Band-Gap (EBG) structures, Frequency Selective Surfaces (FSS), High Impedance Surfaces (HIS), Artificial Magnetic Conductors (AMC), Perfect Magnetic Conductors (PMC), etc. Of all these designations, perhaps the most important are the EBG structures; a comprehensive list of various EBG structures and references, can be found in [273]. In [209] the EBG designation was introduced as a more wide classification to include the others.

All these are in fact artificial impedance surfaces, and can be used to manipulate the propagation of surface waves, control the frequency band (stop, pass, and band gaps), change the surface impedance, control the scattering properties, design tunable impedance surfaces to be used as steerable reflectors or steerable leaky-wave antennas, and many others. This is achieved by modifying the texture or the geometry of the surface, and/or adding other layers; in this way the surface waves and/or the phase of the reflection coefficient of the surface can be manipulated. Concerning the reflection coefficient, the modification of the surface characteristics affects mainly its phase; its magnitude is also affected but less. In specific, a PEC surface causes a 180° shift in the phase of tangential electric component (hence it clearly reflects the field), whereas a PMC lets the phase intact (0° shift); an EBG surface is the most versatile as can vary the phase of the reflected field from $-180°$ to 180°.

Despite their attractive characteristics per case, PEC, PMC and EBG surfaces also exhibit drawbacks when electromagnetic elements that radiate are attached on them; the drawbacks concern the aerodynamic, stealth and conformal properties of the surface. A typical example is the attachment of an electric element on a PEC or a PMC surface.

[46] It is reminded that this is the opposite to a PEC material, in which the *tangential electric* components vanish, and the *normal magnetic* component vanishes.

In the PEC case, attaching the element vertically, its image enhances its radiation and thus the efficiency of the system; however, this configuration has not low profile, something completely undesirable. If this electric element is placed horizontally on the PEC surface to retain low profile, its radiation efficiency decreases significantly because its image introduces a 180° shift and cancels the radiation of the actual element. In the case of a PMC surface, the attachment of an electric element horizontally not only has low profile geometry but also enhances the radiation because its image leaves the phase intact (0° shift). In general, the behavior of electric elements when attached vertically and horizontally on PEC and PMC surfaces is determined by Image Theory [10].

The behavior of EBGs and PMCs when radiating elements are attached on them is similar; however, EBGs have additionally the ability to suppress the surface waves in low profile antennas; microstrip arrays is the most common case[47]. This can reduce the beam scanning capabilities of the microstrip arrays, and in the most extreme case surface waves and coupling may even result to scan blindness.

An EBG surface emulates a PMC surface and suppresses surface waves only over a frequency range; thus, it is usually referred to as a band-gap structure. In general, the frequency band (band-gap) in which an EBG structure operates more efficiently depends upon the application. In next section it is presented how EBG structures, when properly tuned, can be used as PMCs.

5.8.2 PMC realization

Although PMCs do not exist physically, due to their benefits in important applications, it is required to fabricate them artificially; the NRTT structure in the APOTUS-HM is a case where the use of PMCs is indispensable. The last years, PMC surfaces have been synthesized and constructed successfully, and exhibit PMC-type behavior over a frequency range; these surfaces are often called *band-gap* or *band-limited surfaces*. The variety of such surfaces is quite large, and even to simply list them is out of scope of the present text[48]. Next it will be discussed briefly such a surface that can emulate the behavior of PMC. The presentation here is based on [9, 10].

A PMC surface, one of the first fabricated and now widely used, is shown Fig. 5.24. This surface is an array of periodic patches with misc shapes (hexagons here), placed above a very thin substrate (which can be air) and connected to the ground plane; if a non-air substrate is utilized, the connection to the ground plane is done by posts through vias[49], as in usual board circuits. The thickness h of the substrate must be significantly smaller than the operational wavelength λ; usually it is $h < \lambda/10$. The vias are necessary to suppress surface waves within the substrate.

---

[47] In microstrip arrays appear surface waves, which mainly propagate inside the substrate and cause coupling between the array elements.

[48] The interested reader can consult the references on them listed in ch. 8 of [10].

[49] A *via* is an electrical connection between copper layers in a printed circuit board. Essentially a via is a small drilled hole that goes through two or more adjacent layers; the hole is plated with copper that forms electrical connection through the insulation that separates the copper layers (from Wikipedia).

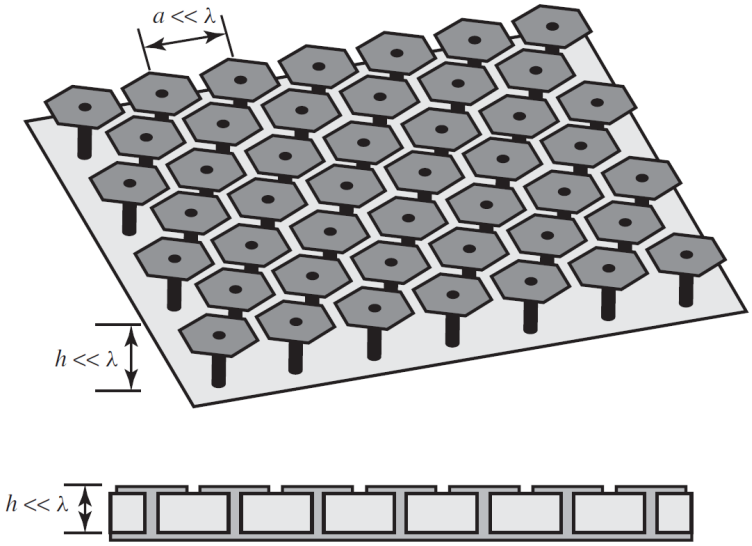


Figure 5.24: (reprinted from [9]). Perspective and side view of a mushroom synthesized surface that can emulate a PMC.

This is an engineered textured surface and belongs to the EBG, PBG and AMC structures mentioned above, and usually it is referred as such. Due to the directional behavior of EBG and PBG structures, when antenna elements are integrated with such structures they can exhibit some remarkable properties [271, 37, 237]. The mushroom EBG surface of Fig. 5.24 can be investigated with a semi-empirical model [9].

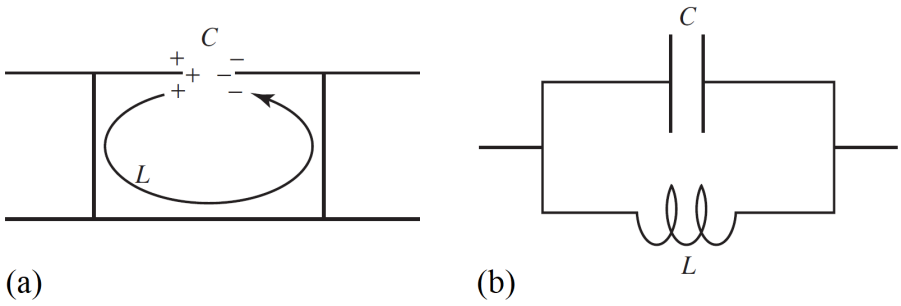


Figure 5.25: (a) Unit cell and (b) equivalent circuit of mushroom textured surface.

In Fig. 5.25a it is shown a unit cell of the EBG structure of Fig. 5.24. An incident wave to an array of such unit cells induces electric fields across the gap of the unit cells; these fields can be represented by an effective capacitance C. Also, the incident waves induce currents that circulate between adjacent unit cells, in paths through the neighboring walls or vias; the effects of these currents can be represented by an equivalent inductance L. Consequently, a model for the unit of Fig. 5.25a is a capacitance C parallel connected with an inductance L, shown in Fig. 5.25b.

With this model, the response of the mushroom surface can be studied. The unit cell of Fig. 5.25b has surface impedance

$$Z_S = i\frac{\omega L}{1-\omega^2 LC}, \tag{5.33}$$

with resonant frequency

$$\omega_0 = \frac{1}{\sqrt{LC}}. \tag{5.34}$$

In practice, the desired resonant frequency is defined by the sheet capacity C_S and the inductance L_S. In their turn, sheet capacity and inductance are defined by the geometry of the individual unit cells and also their arrangement [9]. As can be seen from (5.33), the surface of the unit cell is inductive below the resonant frequency and capacitive above it; near the resonant frequency it becomes very high, and infinity on it. The capacitive or inductive character of the surface impedance defines the type of waves that can be supported: inductive surfaces support TM surface waves, whereas capacitive surfaces support TE surface waves [10]. As reported in [9], this behavior has also been verified experimentally.

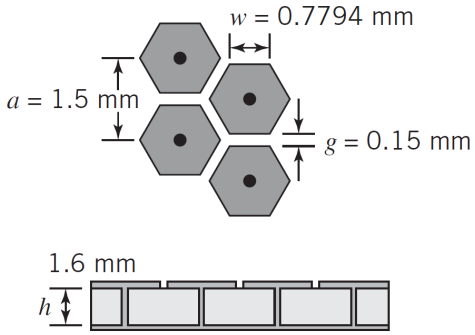


Figure 5.26: (reprinted from [10]). Hexagons in triangular arragement (not shown here), attached on a grounded substrate.

With appropriate choice of geometry and arrangement of the unit cells, the above mushroom surface can exhibit PMC behavior. In Fig. 5.26 it is shown an indicative case. The unit cells (metallic hexagons) are arranged in a triangular lattice, and are attached on the surface of a grounded substrate with electric permitivity 2.2. For the wave excitation, a pair of coaxial probes is placed near the surface; the orientation of these probes controls the excitation of TE or TM modes. In Fig. 5.27 it is shown the amplitude of the transmission between the probes, and also the phase for normal incidence of a plane wave on this surface.

It can be seen in Fig. 5.27a that the herein surface exhibits high impedance in the range about 11-16 GHz (band-gap); below and above the band gap it supports respectively TM surface waves (inductive behavior) and TE surface waves (capacitive behavior).

Concerning the phase response of the surface in the normal incidence of a plane wave, it is evident that the band gap occurs in the range where the phase varies about from −90° to 90°, and it is zero at resonance.

For a textured surface of high impedance like the herein example, the behavior shown in Fig. 5.27 is typical; however, other types or surfaces – and especially those without vertical vias – may exhibit different response characteristics. For the the design of these surfaces (and hence PMCs) a semi-empirical method is discussed in [10].

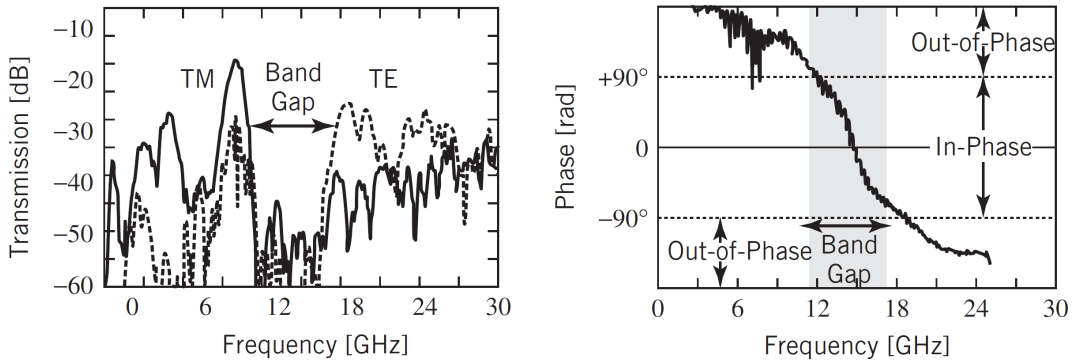


Figure 5.27: (reprinted from [9]). (a) Amplitude of TM (solid) and TE (dashed) modes, and (b) Transmission and phase of a mushroom textured surface with the geometry of Fig. 5.26.

A case where the use of textured surface of high impedance as a PMC can be seen more clearly, is an aperture antenna [9], Fig. 5.28. This antenna has an aperture on a ground plane (reflector) which is of high impedance; here, this surface aims to improve the symmetry of the radiation pattern. The reflector and the unit cells are squares with side length 12.7 cm and 3.7 mm respectively; the antenna is design to work at 12–18 GHz (Ku band). An identical aperture antenna with a metal reflector (i.e., a PEC) is also used for comparison. The radiation patterns for the PEC and the high-impedance surface were measured at 13 GHz, which is within the designed gap; the results are shown in Fig. 5.29. In general, the radiation pattern is affected by the shape and size of the aperture but it is determined mainly by the geometry and texture of the surrounding ground plane (reflector), and the electromagnetic boundary condition of this surface. For the three cases examined here the results are as follows :

(a) Fig. 5.29a. When the ground plane is a conventional metal (PEC), E-plane is usually broader than the H-plane, because for the E-plane its vertical polarized fields are retained by the PEC ground plane, whereas for the H-plane its horizontally polarized fields are, ideally, vanish.

(b) Fig. 5.29b. When the aperture is attached on a textured high-impedance surface and operates within its band gap (here at 13 GHz), its patterns in both the E- and H-planes, are about the same and symmetrical, because the textured surface suppresses both the TM and TE surface waves near the resonant frequency.

(c) Fig. 5.29c. Lastly, when the aperture is attached on the textured surface but operates at the leading edge of the TE band where TM waves are suppressed, then the H-plane pattern is broader than the E-plane; this is the opposite result to Fig. 5.29a concerning the PEC ground plane. This means that the textured ground plane acts as a PMC (ideally eliminates the tangential magnetic fields), which is the opposite of that for the PEC. Therefore, in this case the textured surface emulates a magnetic conductor (PMC).

There is a plethora of applications where textured high-impedance surfaces like the above can be used to control the radiation behavior of the devices; e.g., low-profile antennas, reflective beam steering, leaky wave beam steering, microwave holography are some of them. This functionality is achieved using textured high-impedance surfaces to control, and even eliminate, the surface waves over the band gap and/or the phase of the reflection coefficient – and even make it zero (that is emulate a PMC surface !). For a discussion of the properties EBG surfaces and their applications the reader is referred to [273].

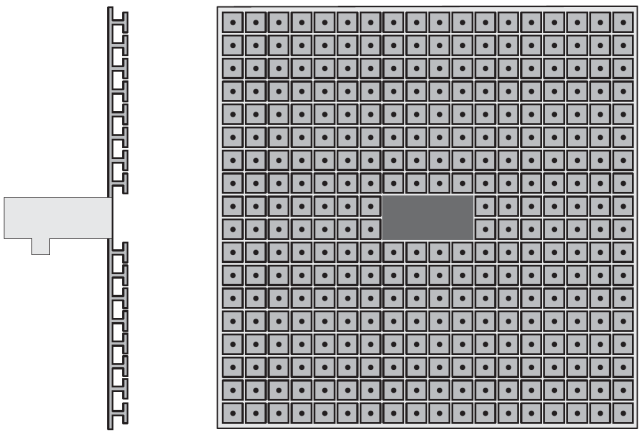


Figure 5.28: (reprinted from [9]). Aperture antenna with a texured surface of high impedance.

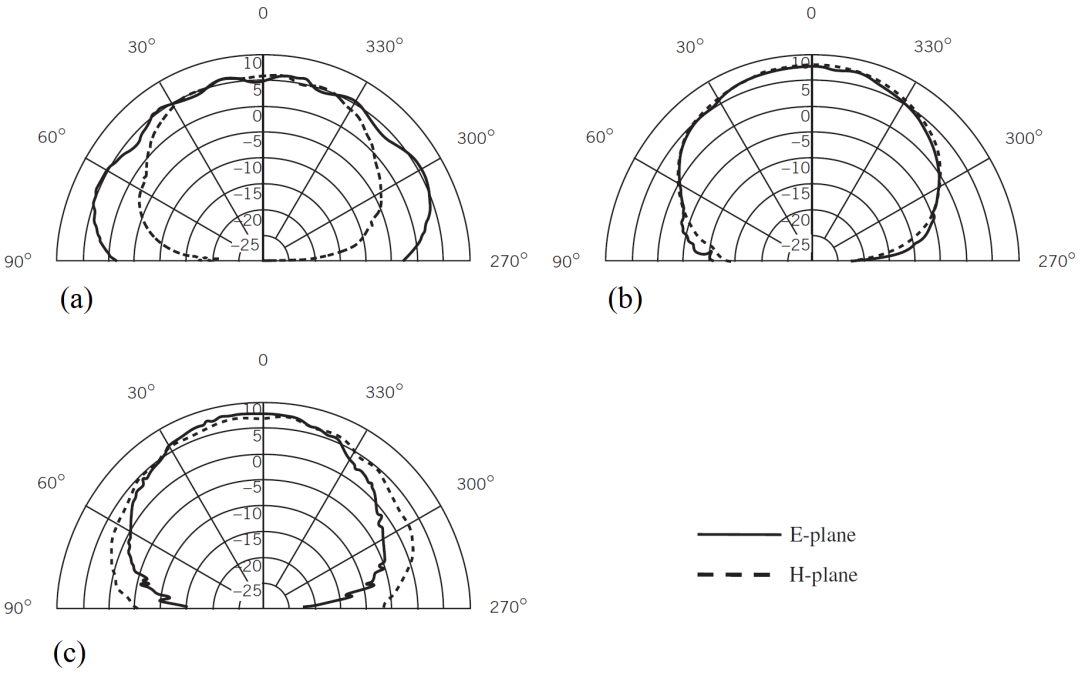


Figure 5.29: (reprinted from [9]). Radiation patterns of an aperture antenna with (a) a conventional ground plane (PEC), (b) a surface of high impedance, (c) a surface of high impedance near the edge of the TE band (PMC).

References

[1] Aharonov Y., Davidovich L., Zagury N., *Phys. Rev.* A **48**, p. 1687, 1993.

[2] Alexandradinata A., Dai X., Bernevig B. A., *Phys. Rev.* B **89**, p. 155114, 2014.

[3] Alpichshev Z. et al., *Phys. Rev. Lett.* **104**, p. 016401, 2010.

[4] Amelio I., Carusotto I., *Phys. Rev.* **101**, p. 064505, 2020.

[5] Ambainis A., *Int. J. Quant. Inf.* **1**, p. 507, 2003.

[6] Anderson P. W., *Phys. Rev.* **109**, p. 1492, 1958.

[7] Asbóth J. K., Oroszlány L., Pályi A., *A Short Course on Topological Insulators*, Springer, 2008.

[8] Bahari B. et al., *Science* **358**, p. 636, 2017.

[9] Balanis C., *Modern Antenna Handbook*, Wiley & Sons 2008.

[10] Balanis C., *Advanced Engineering Electromagnetics 2nd ed.*, Wiley & Sons 2012.

[11] Bandres M. A. et al., *Science* **359**, eaar4005, 2018.

[12] Bandyopadhyay S., Cahay M., *Introduction to Spintronics 2nd ed.*, CRC Press, 2016.

[13] Barnes W. L., Dereux A., Ebbesen T. W., *Nature* **424**, p. 824, 2003.

[14] Baskourelos K. et al., *Phys. Rev. Res.* **4**, p. L032011, 2022.

[15] Belotelov V., Doskolovich L., Zvezdin A., *Phys. Rev. Lett.* **98**, p. 077401, 2007.

[16] Belotelov V. et al., *Nat. Nanotechnol.* **6**, p. 370, 2011.

[17] Bercioux D. et al. (ed.), *Topological Matter. Lectures from the Topological Matter School 2017*, Springer, 2018.

[18] Bernevig B. A., Zhang S. C., *Phys. Rev. Lett.* **96**, p. 106802, 2006.

[19] Bernevig B. A., Hughes T. L., Zhang S. C., *Science* **314**, p. 1757, 2006.

[20] Bernevig B. A., *Topological Insulators and Topological Superconductors*, Princeton University Press, 2013.

[21] Berry M. V., *Proc. R. Soc. Lond.* A **392**, p. 45, 1984.

[22] Bethe H. A., *Phys. Rev.* **66**, p. 163, 1944.

[23] Betzig E., Trautman J. K., *Science* **257**, p. 189, 1992.

[24] Bhargava S., Yablonovitch E., *IEEE Trans. Magn.* **51**, p. 3100407, 2015.

[25] Bhargava S., *Heat-Assisted Magnetic Recording: Fundamental Limits to Inverse Electromagnetic Design*, EECS Department, University of California, Berkeley, Technical Report No. UCB/EECS-2015-106, 2015.

[26] Bittencourt J. A., *Fundamentals of Plasma Physics 3rd ed.*, Springer, 2010.

[27] Blanco-Redondo A. et al., *Science* **362**, p. 568, 2018.

[28] Bohm A. et al., *Int. J. Quantum Chem.* **41**, p. 53, 1992.

[29] Bohren F. C., Huffman R. D., *Absorption and Scattering of Light by small particles*, Willey, 1983.

[30] Bonanni V. et al. *Nano Let.* **11**, p. 5333, 2011.

[31] Bouwkamp C. J., *Philips Research Reports* **5**, p. 321, 1950.

[32] Bouwkamp C. J., *Philips Research Reports* **5**, p. 401, 1950.

[33] Bouwkamp C. J., *Rep. Prog. Phys.* **17**, p. 35, 1954.

[34] Brion J. J. et al., *Phys. Rev. Lett.* **28**, p. 1455, 1972.

[35] Brion J. J., Wallis R. F., *Phys. Rev.* B **10**, p. 3140, 1974.

[36] Broome M. A. et al., *Phys. Rev. Lett.* **184**, p. 153602, 2010.

[37] Brown E. R., Parker D. D., Yablonovitch E., *J. Opt. Soc. Amer.* B **10**, p. 404, 1993.

[38] Brune C. et al., *Phys. Rev. Lett.* **106**, p. 126803, 2011.

[39] Buddhiraju S. et al., *Nat. Commun.* **11**, p. 674, 2020.

[40] Caglayan H., Bulu I., Ozbay E., *Opt. Exp.* **13**, p. 1666, 2005.

[41] Carretero-Palacios S. et al., *Phys. Rev.* B **85**, p. 035417, 2012.

[42] Chang S. H., Gray S. K., Schatz G. C., *Opt. Exp.* **13**, p. 3150, 2005.

[43] Chen C. Y. et al., *Appl. Phys. Lett.* **91**, p. 063108, 2007.

[44] Chen Y. et al., *Science* **325**, p. 178, 2009.

[45] Chen Y. et al., *Phys. Rev. Lett.* **105**, p. 266401, 2010.

[46] Cho J., Angelakis D. G., Bose S., *Phys. Rev. Lett.* **101**, p. 246809, 2008.

[47] Chong Y. D., Wen X. G., Soljačić M., *Phys. Rev.* B **77**, p. 235125, 2008.

[48] Dalibard J. et al., *Rev. Mod. Phys.* **83**, p. 1523, 2011.

[49] de Abajo F. J. G., *Rev. Mod. Phys.* **79**, p. 1267, 2007.

[50] Degiron A., Ebbesen T. W., *Opt. Exp.* **12**, p. 3694, 2004.

[51] Degiron A., et al., *Opt. Comm.* **239**, p. 61, 2004.

[52] Degiron A., Ebbesen T. W., *J. Opt.* A **7**, p. S90, 2005.

[53] Dionne J. et al., *Phys. Rev.* B **73**, p. 035407, 2006.

[54] Drezdzon S. M., Yoshie T., *Opt. Exp.* B **17**, p. 9276, 2009.

[55] Ebbesen T. W. et al., *Nature* **391**, p. 667, 1998.

[56] Ebbesen T. W., Genet C., *Physics Today* **61**, p. 44, 2008.

[57] El-Ganainy R. et al., *Nat. Phys.* **14**, p. 11, 2018.

[58] Eremeev S. V., Koroteeva Y. M., Chulkov E. V., *JETP Lett.* **91**, p. 594, 2010.

[59] Eremeev S. V., Koroteeva Y. M., Chulkov E. V., *JETP Lett.* **92**, p. 161, 2010.

[60] Eschrig H., *Topology and Geometry for Physics*, Springer, 2011.

[61] Fan W. J. et al., *Opt. Exp.* **13**, p. 446, 2005.

[62] Fan W. J. et al., *Phys Rev Lett.* **94**, p. 033902, 2005.

[63] Fang K., Yu Z., Fan S., *Phys. Rev.* B **84**, p. 075477, 2011.

[64] Fang K., Yu Z., Fan S., *Nat. Photonics* **6**, p. 782, 2012.

[65] Fernandes D. E., Silveirinha M. G., *Phys. Rev. Applied* **12**, p. 014021, 2019.

[66] Ferreiro-Vila E. et al., *Phys. Rev.* B **80**, p. 125132, 2009.

[67] Ferreiro-Vila E. et al., *Phys. Rev.* B **83**, p. 205120, 2011.

[68] Frankel T., *The Geometry of Physics 3rd ed.*, Cambridge University Press, 2012.

[69] Fu L., Kane C. L., *Phys. Rev.* B **76**, p. 045302, 2007.

[70] Fu L., Kane C. L., Mele E. J., *Phys. Rev. Lett.* **98**, p. 106803, 2007.

[71] Fu L., *Phys. Rev. Lett.* **103**, p. 266801, 2009.

[72] Gangaraj S. A. H., Monticone F., *Optica* **6**, p. 1158, 2019.

[73] Gangaraj S. A. H., Monticone F., *Phys. Rev. Lett.* **124**, p. 153901, 2020.

[74] Garcia-Vidal F. J. et al., *Phys. Rev. Lett.* **90**, p. 213901, 2003.

[75] Garcia-Vidal F. J. et al., *Appl. Phys. Lett.* **83**, p. 4500, 2003.

[76] Garcia-Vidal F. J. et al., *Phys. Rev. Lett.* **95**, p. 103901, 2005.

[77] Garcia-Vidal F. J. et al., *Rev. Mod. Phys.* **82**, p. 729, 2010.

[78] Genet C., Ebbesen T. W., *Nature* **445**, p. 39, 2007.

[79] Ghaemi H. F. et al., *Phys. Rev.* B **58**, p. 6779, 1998.

[80] Giusfredi G., *Physical Optics. Concepts, Optical Elements and Techniques*, Springer, 2019.

[81] Gong Z. et al., *Phys. Rev.* X **8**, p. 031079, 2018.

[82] Gordon R., Brolo A. G., *Opt. Exp.* **13**, p. 1933, 2005.

[83] Greenstein G., Zajonc A., *The Quantum Challenge: Modern Research on the Foundations of Quantum Mechanics 2nd ed.*, Jones & Bartlett Learning, 2006.

[84] Gresch D. et al., *Phys. Rev.* B **95**, p. 075146, 2017.

[85] Guidry M., Sun Y., *Symmetry, Broken Symmetry, and Topology in Modern Physics*, Cambridge University Press, 2022.

[86] Hadad Y., Steinberg B. , *Phys. Rev. Let.* **105**, p. 233904, 2010.

[87] Hafezi M. et al., *Nat. Phys.* **7**, p. 907, 2011.

[88] Hafezi M. et al., *Nat. Photonics* **7**, p. 1001, 2013.

[89] Hafezi M., Taylor J., *Quantum Simulations with Photons and Polaritons*, Angelakis D. G. (ed.), Springer, 2017.

[90] Haldane F. D. M., *Phys. Rev. Lett.* **61**, p. 2015, 1988.

[91] Haldane F. D. M., Raghu S., *Phys. Rev. Lett.* **100**, p. 013904, 2008.

[92] Halperin B. I., *Phys. Rev.* B **25**, p. 2185, 1982.

[93] Han C. et al., *Light Sci. Appl.* **8**, p. 40, 2019.

[94] Harari G. et al., *Science* **359**, eaar4005, 2018.

[95] Hasan M. Z., Kane C. L., *Rev. Mod. Phys.* **82**, p. 3045, 2010.

[96] Haus H. A., *Waves and Fields in Optoelectronics*, Prentice Hall, 1984.

[97] Hibbins A. P., Evans B. R., Sambles J. R., *Science* **308**, p. 670, 2005.

[98] Hodgson N., Weber H., *Laser Resonators and Beam Propagation 2nd ed.*, Springer, 2005.

[99] Hsieh D. et al., *Nature* **452**, p. 970, 2008.

[100] Hsieh D. et al., *Science* **23**, p. 919, 2009.

[101] Hu B. et al., *Opt. Comm.* **24**, p. 6120, 2008.

[102] Hu B., *Plasmonics*, Gric T. (ed.), ch. 6, 2018.

[103] Ibach H., Luth, H., *Solid-State Physics. An Introduction to Principles of Materials Science 4th ed.*, Springer, 2009.

[104] Jackiw R., Rebbi C., *Phys. Rev.* D **13**, p. 3398, 1976.

[105] Jaksch D., Zoller P., *New J. Phys.* **5**, p. 56, 2003.

[106] Jackson J. D., *Classical Electrodynamics 3rd ed.*, Willey, 1999.

[107] Jia N. et al., arXiv preprint at http://lanl.arxiv.org/abs/1309.0878, 2013.

[108] Jin D. et al. *Nat. Commun.* **7**, p. 13486, 2016.

[109] Jin H. et al. *Phys. Rev.* B **83**, p. 041202, 2011.

[110] Joannopoulos J. D., Johnson S. G., Winn J. N., Meade R. D., *Photonic Crystals. Molding the Flow of Light 2nd ed*, Princeton University Press, 2008.

[111] Johnson B. L., Shiau H. H., *J. Phys.: Cond. Mat.* **20**, p. 335217, 2008.

[112] Joushaghani A. et al., *Opt. Exp.* **19**, p. 8367, 2011.

[113] Jung J. et al., *Phys. Rev.* B **79**, p. 153407, 2009.

[114] Jung J., Martin-Moreno L., Garcia-Vidal F. J., *New J. Phys.* **11**, p. 123013, 2009.

[115] Kalajdzievski S., *An illustrated introduction to Topology and Homotopy*, CRC Press, 2015.

[116] Kane C. L., Mele E. J., *Phys. Rev. Lett.* **95**, p. 226801, 2005.

[117] Kane C. L., Mele E. J., *Phys. Rev. Lett.* **95**, p. 146802, 2005.

[118] Kempe J., *Contemp. Phys.* **44**, p. 307, 2003.

[119] Khanikaev A. B. et al., *Phys. Rev. Let.* **105**, p. 126804, 2010.

[120] Khanikaev A. B. et al., *Nat. Mater.* **12**, p. 233, 2013.

[121] Khanikaev A. B., Shvets G., *Nat. Photonics* **11**, p. 763, 2017.

[122] Kim D. et al., *Nat. Commun.* **4**, p. 2040, 2013.

[123] Kim J. H., Moyer P. J., *Opt Exp.* **14**, p. 6595, 2006.

[124] Kinsey L. C., *Topology of Surfaces*, Springer, 1993.

[125] Kitaev A., *AIP Conf. Proc.* **1134**, p. 22, 2009.

[126] Kitagawa T., *Quantum Inf. Process* **11** p. 1107, 2012.

[127] Kitagawa T. et al., *Nat. Comm.* **3**, p. 882, 2012.

[128] Klafter J., Sokolov M., *First Steps in Random Walks*, Oxford University Press, 2011.

[129] v. Klitzing K., Dorda G., Pepper M., *Phys. Rev. Lett.* **45**, p. 494, 1980.

[130] Koch J. et al., *Phys. Rev.* A **82**, p. 043811, 2010.

[131] Koerkamp K. J. et al., *Phys. Rev. Lett.* **92**, p. 18390, 2004.

[132] Kong F. et al., *Prog. Elect. Res.* **82**, p. 257, 2008.

[133] Konig M. et al., *Science* **318**, p. 766, 2007.

[134] Konig M. et al., *J. Phys. Soc. Jpn.* **77**, p. 031007, 2008.

[135] Kraus E. Y. et al., *Phys. Rev. Lett.* **109**, p. 106402, 2012.

[136] Krishnan A. et al., *Opt. Comm.* **200**, p. 1, 2001.

[137] Kuroda K. et al., *Phys. Rev. Lett.* **105**, p. 146801, 2010.

[138] Kuroda K. et al., *Phys. Rev. Lett.* **105**, p. 076802, 2010.

[139] Kuroda K. et al., *Phys. Rev. Lett.* **108**, p. 206803, 2012.

[140] Kurokawa Y., Miyazaki H., *Phys. Rev.* B **75**, p. 035411, 2007.

[141] Kushwaha M. S., *Surf. Sc. Rep.* **41**, p. 1, 2001.

[142] Lalanne P. et al., *Surface Sci. Rep.* **64**, p. 453, 2009.

[143] Lan Y. C., Chang Y. C., Lee P. H., *Appl. Phys. Lett.* **90**, p. 171114, 2007.

[144] Lan Y. C., Chen C. M., *Opt. Exp.* **18**, p. 12470, 2010.

[145] Lee T. E., *Phys. Rev. Lett.* **116**, p. 133903, 2016.

[146] Leone R., *J. Phys.* A **44**, p. 295301, 2011.

[147] Leviatan Y., Harrington R. F., Mautz J. R., *IEEE Trans. Antennas Propag.* **30**, p. 1153, 1182.

[148] Lesuffleur A., Kumar L, Gordon R., *Appl. Phys. Lett.* **88**, p. 261104, 2006.

[149] Leykam D. et al., *Phys. Rev. Lett.* **118**, p. 040401, 2017.

[150] Li Y. Y. et al., *Adv. Mater.* **22**, p. 4002, 2010.

[151] Li Y., *Plasmonic Optics. Theory and Applications*, SPIE, 2017.

[152] Lin H. et al., *Phys. Rev. Lett.* **105**, p. 036404, 2010.

[153] Liu C. X. et al., *Phys. Rev.* B **82**, p. 045122, 2010.

[154] Liu H. T., Lalanne P., *Nature* **452**, p. 728, 2009.

[155] Liu K., Shen L., He S., *Opt. Lett.* B **37**, p. 4110, 2011.

[156] Liu X. X. et al., *Appl. Opt.* **48**, p. 3102, 2009.

[157] Liu K., Shen L., He S., *Novel In-Plane Semiconductor Lasers XIX*, A. A. Belyanin and P. M. Smowton (ed.), SPIE, 2020

[158] Liu Z., Jin G., *Appl. Phys.* A **105**, p. 819, 2011.

[159] Longhi S., *Europhys. Lett.* **120**, p. 64001, 2017.

[160] Lu L., Joannopoulos J. D. Soljačić M., *Nat. Photonics* **7**, p. 294, 2013.

[161] Lu L., Joannopoulos J. D. Soljačić M., *Nat. Photonics* **8**, p. 821, 2014.

[162] MacDonald A. H., Streda P., *Phys. Rev.* B **29**, p. 1616, 1984.

[163] Maier S. A., *Plasmonics. Fundamentals and applications*, Springer, 2007.

[164] Maradudin A. A., *Structured Surfaces as Optical Metamaterials*, Cambridge University Press, 2011.

[165] Mary A. et al., *Phys. Rev.* B **80**, p. 165431, 2009.

[166] Martin-Moreno L. et al., *Phys. Rev. Lett.* **86**, p. 1114, 2001.

[167] Martin-Moreno L., Garcia-Vidal F. J., Pendry J. B., *J. Opt. A: Pure Appl. Opt.* **7**, p. S97, 2005.

[168] Moreno E., Martin-Moreno L., Garcia-Vidal F. J., *J. Opt. A: Pure Appl. Opt.* **8**, p. S94, 2006.

[169] Marzari N., Vanderbilt D., *Phys. Rev.* B **56**, p. 12847, 1997.

[170] Mermin N. D ., *Quantum Computer Science: An Introduction*, Cambridge University Press, 2007.

[171] Mittal S. et al., *Phys. Rev. Lett.* **113**, p. 087403, 2014.

[172] Mittal S., Orre V. V., Hafezi M., *Opt. Exp.* **24**, p. 15631, 2016.

[173] Moore J. E., Balents L., *Phys. Rev.* B **75**, p. 121306(R), 2007.

[174] Nakahara M., *Geometry, Topology and Physics*, CRC Press, 2003.

[175] Nielsen M. A., Chuang I. L., *Quantum Computation and Quantum Information*, Cambridge University Press, 2000.

[176] Nikitin A. Yu., Garcia-Vidal F. J., Martin-Moreno L., *J. Opt. A: Pure Appl. Opt.* **11**, p. 125702, 2009.

[177] Nolting W., *Theoretical Physics vol. 6. Quantum Mechanics - Basics*, Springer, 2017.

[178] Nolting W., *Theoretical Physics vol. 7. Quantum Mechanics - Methods and Applications*, Springer, 2017.

[179] Novotny L., Hecht B., *Principles of Nano-Optics 2nd ed.*, Cambridge University Press, 2012.

[180] Orbons S. M., Roberts A., *Opt. Exp.* **14**, p. 12623, 2006.

[181] Orbons S. M. et al. *Appl. Phys. Lett.* **90**, p. 251107, 2007.

[182] Orbons S. M. et al., *Opt. Lett.* **33**, p. 821, 2008.

[183] Ota Y. et al., *Nanophotonics* **9**, p. 547, 2020.

[184] Ozawa T. et al., *Rev. Mod. Phys.* **91**, p. 015006, 2019.

[185] Ozbay E., *Science* **311**, p. 189, 2006.

[186] Palik E. D., Furdyna J. K., *Rep. Prog. Phys.* **33**, p. 1193, 1970.

[187] Palik E. D. et al., *Phys. Rev.* B **13**, p. 2497, 1976.

[188] Pancharatnam S., *Proc. of the Indian Academy of Sciences* A **44**, p. 247, 1956.

[189] Pankratov O. A., Pakhomov S. V., Volkov B. A., *Sol. State Commun.* **61**, p. 93, 1987.

[190] Pilozzi L., Conti C., *Phys. Rev.* B **93**, p. 195317, 2016.

[191] Popov E. et al., *Appl. Opt.* **44**, p. 2332, 2005.

[192] Porto J. A., Garcia-Vidal F. J., Pendry J. B., *Phys. Rev. Lett.*, **83**, p. 2845, 1999.

[193] Portugal R., *Quantum Walks and Search Algorithms 2nd ed.*, Springer, 2018.

[194] Prikulis J. et al., *Nano Lett.* **4**, p. 1003, 2004.

[195] Przybilla F. et al., *J. Opt. A, Pure Appl. Opt.* **8**, p. 458, 2008.

[196] Przybilla F. et al., *Opt. Exp.* **16**, p. 9571, 2008.

[197] Puentes G., *Crystals* **7**, p. 122, 2017.

[198] Qi X. L., Zhang S. C., *Rev. Mod. Phys.* **83**, p. 1057, 2011.

[199] Raether H., *Surface Plasmons on Smooth and Rough Surfaces and on Gratings*, Springer, 1988.

[200] Raghu S., Haldane F. D. M., *Phys. Rev.* A **78**, p. 033834, 2008.

[201] Raza S. et al., *J. Phys.: Cond. Mat.* **27**, p. 183204, 2015.

[202] Rayleigh L., *Philos. Mag.* **14**, p. 60, 1907.

[203] Rechtsman M. C. et al., *Nature* **496**, p. 196, 2013.

[204] Rechtsman M. C. et al., *Optica* **3**, p. 925, 2013.

[205] Reif F., *Fundamentals of Statistical and Thermal Physics*, Waveland Pr. Inc., 1965.

[206] Rodrigo S. G., Garcia Vidal F. J., Martin-Moreno L., *Phys. Rev.* B **77**, p. 075401, 2008.

[207] Roth A. et al., *Science* **325**, p. 294, 2009.

[208] Rudner M. S., Levin M., Levitov L. S., *Cond. Mat.* **325**, arXiv:1605.07652, 2016.

[209] Rahmat-Samii Y., Mosallaei H., *Electromagnetic Band-Gap Structures: Classification, Characterization, and Applications*, 11th International Conference on Antennas and Propagation (ICAP 2001), Manchester, UK, p. 564, April 17–20, 2001.

[210] Sakurai J. J., Napolitano J., *Modern Quantum Mechanics 3rd ed.*, Cambridge University Press, 2021.

[211] Salomon L. et al., *Phys. Rev. Lett.* **86**, p. 1110, 2001.

[212] Sato T. et al., *Phys. Rev. Lett.* **105**, p. 136802, 2010.

[213] Schmidt, M., Peano, V., Marquardt F., arXiv preprint at http://lanl.arxiv.org/abs/1311.7095, 2013.

[214] Segev M., Bandres M. A., *Nanophotonics* **10**, p. 425, 21.

[215] Sgiarovello C., Peressi M., Resta R., *Phys. Rev.* B **64**, p. 115202, 2001.

[216] Shadrivov I. V., Sukhorukov A. A., Kivshar Y. S., *Phys. Rev.* E **67**, p. 057602, 2003.

[217] Shen H., Zhen B. Fu L., *Phys. Rev.* **120**, p. 146402, 2018.

[218] Shen L. et al., *Opt. Exp.* **23**, p. 950, 2015.

[219] Shi T., Kimble H. J., Cirac J. I., *Proc. Natl. Acad. Sci. USA* **114**, p. E8967, 2017.

[220] Shin H., Catrysse P. B., Fan S., *Phys. Rev.* B **72**, p. 085436, 2005.

[221] Silveirinha M., Engheta N., *Phys. Rev. Lett.* **97**, p. 157403, 2006.

[222] Silveirinha M., *Phys. Rev. Lett.* B **92**, p. 125153, 2015.

[223] Silveirinha M., *Phys. Rev.* B **94**, p. 205105, 2016.

[224] Silveirinha M., *Phys. Rev.* B **97**, p. 115146, 2018.

[225] Silveirinha M., *Phys. Rev.* X **9**, p. 011037, 2019.

[226] Simon D. S. et al., *Phys. Rev.* A **96**, p. 013858, 2017.

[227] Simon D. S., *Topology in Optics: Tying Light in Knots*, IOP Publishing, 2022.

[228] Skirlo S. A., Lu L., Soljačić M., *Phys. Rev. Lett.* **113**, p. 113904, 2014.

[229] Schnyder A. P. et al., *Phys. Rev.* B **78**, p. 195125, 2008.

[230] Shao Z. K. et al., *Nat. Nanotechnol.* **15**, p. 67, 2020.

[231] Soluyanov A. A., Vanderbilt D., *Phys. Rev.* B **83**, p. 235401, 2011.

[232] Soluyanov A. A., Vanderbilt D., *Phys. Rev.* B **83**, p. 035108, 2011.

[233] Soluyanov A. A., *Topological Aspects of Band Theory*, PhD thesis, Rutgers University - Graduate School - New Brunswick, 2012.

[234] Soluyanov A. A., Vanderbilt D., *Phys. Rev.* B **85**, p. 115415, 2012.

[235] Souma S. et al., *Phys. Rev. Lett.* **108**, p. 116801, 2012.

[236] Sun M. et al., *Phys. Rev. Lett.* A **365**, p. 512, 2012.

[237] Suzuki T., Yu L., *J. Appl. Phys.* A **49**, p. 582, 1996.

[238] Su W. P., Schrieffer J. R., Heeger A. J., *Phys. Rev. Lett.* **42**, p. 1698, 1979.

[239] Tarasinski B., Asbóth J. K., Dahlhaus J. P., *Phys. Rev.* A **89**, p. 042327, 2014.

[240] Temnov V. et al., *Nat. Photonics* B **4**, p. 107, 2010.

[241] Thonhauser T., Vanderbilt D., *Phys. Rev.* B **73**, p. 235111, 2006.

[242] Thouless D. J. et al., *Phys. Rev. Lett.* **49**, p. 405, 1982.

[243] Thouless D. J. et al., *Phys. Rev.* B **27**, p. 405, 1983.

[244] Thouless D. J., *J. Phys.* C **17**, p. L325, 1984.

[245] Tkachov G., *Topological Insulators. The Physics of Spin Helicity in Quantum Transport*, Pan Stanford Publishing, 2015.

[246] Torrado J. F. et al., *Opt. Exp.* **18**, p. 15635, 2010.

[247] Tsakmakidis K. L. et al., *Phys. Rev.* B **73**, 2006.

[248] Tsakmakidis K. L. et al., *Science*, **356**, 2017.

[249] Tsakmakidis K. L. et al., *Science*, **358**, eaan5196, 2017.

[250] Tsakmakidis K. L. et al., *Optica*, **7**, p. 1097, 2020.

[251] Tsakmakidis K. L., Baskourelos K., Stefanski T., *Appl. Phys. Lett.*, **119**, p. 190501, 2021.

[252] Vanderbilt D., *Berry Phases in Electronic Structure Theory*, Cambridge University Press, 2018.

[253] Van der Molen K. L. et al., *Phys. Rev.* B **72**, p. 045421, 2005.

[254] Venegas-Andraca S. E., *Quantum Walks for Computer Scientists*, Morgan and Claypool Publishers, 2008

[255] Venegas-Andraca S. E., *Quantum Inf. Process.* **11**, p. 1015, 2012.

[256] Visser T. D., *Nature Phys.* **2**, p. 509, 2006.

[257] Volkov B. A., Pankratov O. A., *JETP Lett.* **42**, p. 178, 1985.

[258] Wallis R. F., Brion J. J., Burstein E., *Phys. Rev.* B **9**, p. 3140, 1974.

[259] Wang J., Manouchehri K., *Physical Implementation of Quantum Walks*, Springer, 2014.

[260] Wang J. et al., *Phys. Lett.* A, **398**, p. 127279, 2021.

[261] Wang M. et al., *Nanophotonics* **8**, p. 1327, 2019.

[262] Wang Z. et al., *Phys. Rev. Lett.* **100**, p. 013905, 2008.

[263] Wang Z. et al., *Nature* **461**, p. 772, 2009.

[264] Webb K. J., Li J., *Phys. Rev.* B **73**, p. 033401, 2006.

[265] Winnewisser C. et al., *Appl. Opt.* **38**, p. 3961, 1999.

[266] Wilczek F., Shapere A., *Geometric Phases in Physics. Vol. 5*, World Scientific, 1989.

[267] Winkler G. W., Soluyanov A. A., Troyer M., *Phys. Rev.* B **93**, p. 035453, 2016.

[268] Wu T. K., Lee S. W., *IEEE Trans. Antennas Propag.* **42**, p. 1484, 1994.

[269] Xiao L. et al., *Nat. Phys.* **16**, p. 761, 2020.

[270] Xia Y. et al., *Nat. Phys.* **5**, p. 398, 2009.

[271] Yablonovitch E., *J. Opt. Soc. Amer.* B **10**, p. 283, 1993.

[272] Yan B. et al., *EPL* **90**, p. 37002, 2010.

[273] Yang F., Rahmat-Samii Y., *Electromagnetic Band Gap Structures in Antenna Engineering*, Cambridge University Press, 2008.

[274] Ye Y. H. et al., *Appl. Phys. Lett.* **91**, p. 25105, 2007.

[275] Yi Ju-Min et al., *ACS Photonics* **1**, p. 365, 2014.

[276] Yu Z. et al., *Phys. Rev. Lett.* **100**, p. 23902, 2008.

[277] Zak J., *Phys. Rev. Lett.* **62**, p. 2747, 1989.

[278] Zeng Y. et al., *Nature* **578**, p. 246, 2020.

[279] Zeuner J. M. et al., *Phys. Rev. Lett.* **115**, p. 040402, 2015.

[280] Zhang H. et al., *Nat. Phys. Rev.* **5**, p. 438, 2009.

[281] Zimmermann S. T., *Nanoscale Lithography and Thermometry with Thermal Scanning Probes*, Thesis No 9060,
École Polytechnique Fédérale de Lausanne, Switzerland, 2018.

[282] Zhu H., Jiang C., *Opt. Lett.* **36**, p. 1308, 2011.

Books/Monographs :

[283] Tsakmakidis K. L., Baskourelos K., Wartak M.,
Metamaterials and Nanophotonics; Principles, Techniques, and Applications,
World Scientific, 2022.

Publications in Archival Journals and Magazines :

[284] Xu J., Luo Y., Yong K., Baskourelos K., Tsakmakidis K. L.,
"Extraordinary optical transmission from a guided mode to free-space radiation",
Comm. Phys. 2023, (Nature Publishing Group; under 2nd round of review).

[285] Baskourelos K. et al.,
"Topological extraordinary optical transmission",
Phys. Rev. Research **4**, L032011, 2022,
DOI: https://doi.org/10.1103/PhysRevResearch.4.L032011.

[286] Tsakmakidis K. L., Baskourelos K., Stefanski T.,
"Topological, nonreciprocal, and multiresonant slow light beyond the time-bandwidth limit",
Appl. Phys. Lett. **119**, 190501, 2021,
(invited Perspectives article, also selected as Featured),
DOI: 10.1063/5.0068285.

[287] Stefański T., Baskourelos K., Tsakmakidis K.,
"Finite-difference time-domain analyses of active cloaking for electrically-large objects",
Opt. Exp. **29**, p. 355, 2021.

[288] Zouros G., Kolezas G., Almpanis E., Baskourelos K., Stefański T., Tsakmakidis K.,
"Magnetic switching of Kerker scattering in spherical microresonators",
Nanophotonics, 2020,
DOI: https://doi.org/10.1515/nanoph-2020-0223.

International Conferences and Topical Meetings :

[289] Baskourelos K., Tsakmakidis K.,
"Science and applications of topological rainbow trapping",
Metamaterials' 2023 Conference (11 – 16 Sept. 2023, Crete, Greece). (invited)

[290] Baskourelos K., Tsakmakidis K.,
"Topological trapped-rainbow and nonreciprocal guides beyond the time-bandwidth limit",
META 2023 Conference (18 – 21 July 2023, Paris, France). (invited)

[291] Baskourelos K., Tsakmakidis K.,
"Topological 'perfect' focusing and giant local-field enhancements",
in 2023 Optica Nonlinear Optics Topical Meeting (10 – 13 July 2023, Honolulu, Hawaii, USA).

[292] Baskourelos K., Tsakmakidis K.,
"Topological trapped-rainbow and nonreciprocal guides beyond the time-bandwidth limit",
in 2022 International Workshop on "Structured materials and structured light," Ettore Majorana Foundation, Erice (Sicily), Italy.
(invited by Prof. Federico Capasso, Harvard Univ.). (invited)

[293] Baskourelos K., Tsakmakidis K.,
"Topological slow light beyond the time-bandwidth limit",
in Novel Concepts and New Materials, Electromagnetic Metamaterials and Metasurfaces: Recent Research Achievements and New Paradigms, 9th Forum on New Materials, CIMTEC 2022 (Perugia, Italy, 25–29 June 2022),
paper FF-2:IL03. (invited)